Machine Learning Approaches to Assessing Future Flood & Storm Risk

Robert Edwin Rouse

Department of Engineering
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Darwin College

January 2023

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Robert Edwin Rouse
January 2023

Acknowledgements

First and foremost, I would like to thank my supervisors Professor Allan McRobie, Professor Emily Shuckburgh, and Dr Scott Hosking; without their help and guidance this work would have been far lesser, if not impossible. They helped shape the narrative arc, pruned back the ideas that were not viable, and gave impetus and ambition to the ones that were.

Secondly, I would like to thank my industrial supervisors, Professor David Viner and Dr Sun Yan Evans, who highlighted how to make research impactful in a real world context and how to make it transferable.

Professor Mark Girolami, Will Tebbutt, Risa Ueno, Marc Girona-Mata, Rachel Furner, Tom Andersson, and John Bronskill provided much needed sanity checks throughout my doctorate; the countless discussions that I had with each of them were enjoyable and thought provoking and the wider research group as a whole was also a source of inspiration.

My penultimate thanks goes to my family, who have always supported my research and studies and provided both a respite from my academic life at Cambridge and advice when needed.

Finally, I would like to acknowledge my partner, Phoebe, for her eternal patience and love and without whom I would not be the person I am today, nor would I have come close to the finish line.

My doctorate was funded by a studentship in the Future Infrastructure and Build Environment Centre for Doctoral Training from the Engineering and Physical Sciences Research Council; an Industrial Fellowship from the Royal Commission for the Great Exhibition of 1851, in conjunction with Mott MacDonald; and in affiliation with the Artificial Intelligence for Environmental Risk Centre for Doctoral Training. The support of all these institutions made my research possible.

Machine Learning Approaches to Assessing Future Flood & Storm Risk

Robert Edwin Rouse

Abstract

This thesis describes the application of machine learning to hydrology problems in the face of imminent and long term climate change, in particular through the lens of data minimalism. First, we note that with the dawn of the Anthropocene the world's climate is changing, primarily due to human activity. That change is having and will continue to have profound effects on climatic, geological, and biological systems, including the world's hydrological systems with which we are concerned. We further note that there are large swathes of the world where there is insufficient data for the development of traditional empirical models, so a new approach is required, one that can generalise from the data we do have.

The opening chapters begin with a treatment of the relevant features and a strategy for compressing the dimensionality to create a model that does not rely on internal measurements from a hydrological system to make predictions about streamflow. We then take this framework and examine the performance of different machine learning model architectures within it and note that whilst some machine learning methods offer greater flexibility, simplicity, and performance, this is subordinate to the feature engineering.

With the modelling approach established for a single hydrological system, we extend the framework in order to be able to better generalise to unseen hydrological systems; by introducing additional features that describe the physical nature of the catchment, such as its topography and geology and again noting that these might be assessed externally, along with proxy variables to estimate anthropogenic interaction with the natural system. In doing so, we create a framework that is able to generate usable predictions for even the most problematic of systems, such as those where human activity has resulted in severe ecosystem degradation.

If flooding is of concern then so too should be extreme phenomena that result in the causal precipitative events; and thus our attention turns toward the application of machine learning to storm prediction. Our experimental approach offers a brief guide through a history of networks for computer vision tasks, as any gridded data is analogous to an image, before developing a multi-task approach for identifying cyclonic activity, locating it within a domain, and then predicting its likely precipitative impact, thus connecting it to flood risk.

In attempting to improve certain frameworks for their better application to the problems at hand, namely those concerned with time histories, antecedent dependencies, or hidden states, we modify and extend the Neural Process framework, examining the structure of the encoder and decoder networks within. Whilst we achieve state of the art performance with the first foray, using Recurrent Neural Networks as encoder and decoder, our second, using Temporal Convolutional Neural Networks, was less successful but does not diminish our belief in the viability of the general approach.

Finally, we examine the application of the methods described in this thesis to the problem of future scenario impact modelling using climate projections; although ubiquitous methods of correcting the systematic bias within climate models were used, we believe that a new approach is required, one where climate projections are used to force a generative machine learning weather model that is then used to force impact models. This approach has the potential to be more realistic and more powerful, rapidly generating thousands of scenarios of impact under a given climate projection to create informed statistical distributions of impact.

Table of Contents

List of Figures

List of Tables

Nomenclature

Latin Symbols

A	Catchment area
C	Coordinate, either longitudinal or latitudinal
E	Evapotranspiration
H	Internal Storage
J	Cost function
K	Kernel matrix
P	Probability, either distribution or of a variable
Q	Quantile
S	Sensitivity
T	Time period length
U	Probability distribution
V	Probability distribution
W	Wasserstein Distance
X	Input matrix
Y	Output matrix
a	Interim output
c	Coefficient or constant

d	Dimension or dimensional coefficient
e	Euler's number, ≈ 2.71828
h	Hidden state
k	Kernel function
l	Layer or layer number
n	Node, nodal number, or number of data points
r	Representation encoding
t	Time step
u	u component of wind speed
v	v component of wind speed
w	w component of wind speed
x	Input variable
y	Output variable
z	Parameterisation

Greek Symbols

Γ	Gating function
Δ	Large change in
Θ	Set or tensor of hyperparameters
Λ	Wind speed vector
Ξ	Scoring criterion pertaining to subscript
Π	Set of couplings
Φ	Tensor, matrix, or column of weights
Ψ	Precipitation
Ω	Climatic variable or set of variables

α	Activation function
β	Bias or constant
γ	Relevance weighting function
δ	Small change in
ε	Error or adjustment term
ζ	Relative vorticity
η	Learning rate
θ	Hyperparameter
κ	Scaling factor or vector
λ	Lengthscale
μ	Mean or mean function
ν	Variance
π	Coupling
ρ	Percentile
σ	Standard deviation
ϕ	Weighting parameter
ψ	Variance
ω	Vorticity

Superscripts

i	Superscript index
j	Superscript index
T	Matrix transpose
$'$	Predictand
$*$	Variable to the test set or at a point to which a function is fit

$\hat{}$	Unit vector

Subscripts

a	Accuracy score
d	Depth
f	F1 score
h	Height
i	Subscript index
j	Subscript index
n	Iterative maximum
p	Precision
r	Recall
w	Width
$*$	Matrix or function of evaluations using training and test set values
$**$	Matrix or function of evaluations using test set values

Other Symbols

$\mathscr{D}$	Dataset of observations
$\mathscr{GP}$	Gaussian process distribution
∇	Vector differential operator
$\mathscr{N}$	Normal distribution
$\mathscr{O}$	Computational complexity
$\mathscr{Y}$	Output domain
∇	Vector differential operator
$\bigodot$	Sequence of tensor convolutions
$\odot$	Tensor convolution

Acronyms / Abbreviations

ANN	Artificial Neural Network
BCE	Binary Cross Entropy
CMIP6	Coupled Model Intercomparison Project Phase 6
CNN	Convolutional Neural Method
ERA	Environmental Reanalysis
FN	False Negative
FP	False Positive
GCM	General Circulation Model
GP	Gaussian Process
GRU	Gated Recurrent Unit
IM-**H**	Impact Model Hindcast-trained
IM-**O**	Impact Model Observation-trained
IPCC	Intergovernmental Panel on Climate Change
KLD	Kullback-Liebler Divergence
LSTM	Long Short-Term Memory
MLP	Multi-layer Perceptron
MPI-ESM	Max Planck Institute Earth System Model
MSE	Mean Squared Error
NLL	Negative Loss Likelihood
NP	Neural Process
NSE	Nash-Sutcliffe Efficiency
NSERA	Nash-Sutcliffe Efficiency Relevance Area
RCM	Regional Circulation Model

RNN	Recurrent Neural Method
RNP	Recurrent Neural Process
SSP	Shared Socioeconomic Pathway
SSVGP	Sparse Stochastic Variational Gaussian Process
TCNN	Temporal Convolutional Neural Network
TNP	Temporal Neural Process
TN	True Negative
TP	True Positive
WD	Wasserstein Distance

Chapter 1

Introduction

1.1 A Changing World

The world's climate is changing. Whilst there are variations that have occurred throughout history, the nature of this change at the end of the Holocene has been driven primarily by human activity; with the extent of those changes being significant enough to alter climatic, geological, and biological signals, this period has even warranted its own moniker, the Anthropocene (Waters et al., 2016).

Taking climate alone, there has been an upward trend in global atmospheric carbon dioxide data since the time of the Industrial Revolution, compiled using adjusted data from a combination of ice cores, for the years 1850-1958, and direct measurements, for the years 1959-2020, (Etheridge et al., 1996; Tans and Keeling, 2022); the mean annual global concentration of atmospheric carbon dioxide is estimated to have increased by a factor of approximately 1.5, rising from approximately 285 parts per million to over 415 parts per million, as shown in Figure 1.1.

The altered atmospheric composition, through carbon dioxide and other Greenhouse Gases (GHGs), has implications for the Earth's energy balance, with regards to the amount of predominantly solar radiation that is either absorbed or reflected. Radiative forcing is the change in that balance, expressed as the change in the amount of energy absorbed per unit area, with a positive forcing arising from an increase over the pre-industrial era of approximately $1.5Wm^{-2}$, far above the $1Wm^{-2}$ recommended limit for avoiding irreversible climate change (Rockström et al., 2009); corresponding with that increased radiative forcing, there has been a rise in average surface air temperature of more than 1 °C since 1850, where a safe limit to avert irreversible climate change has been recommended at 1.7 °C.

The mounting evidence for climate change driven by anthropogenic emissions of greenhouse gases is incontrovertible with various "tipping points", serving great importance to

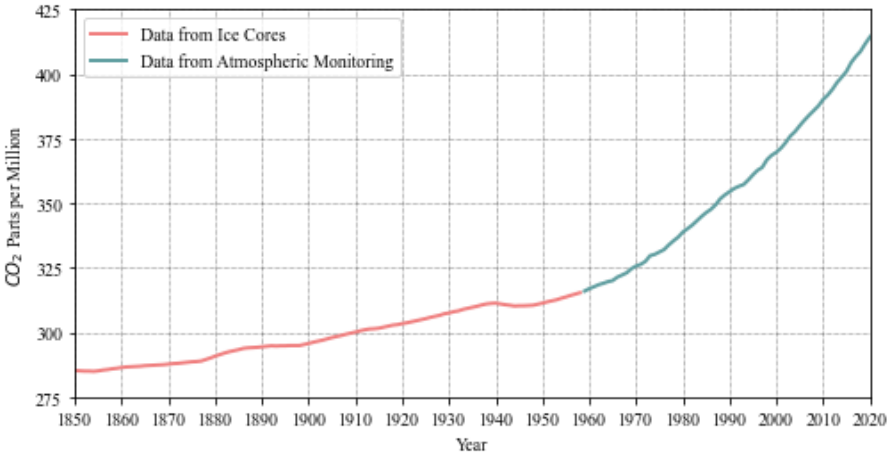


Fig. 1.1 Mean annual global concentration of atmospheric carbon dioxide using adjusted data from ice cores and direct measurements

either Earth system functioning or contributing to human welfare and society, close to being passed (McKay et al., 2022), such as the collapse of the Greenland and West Antarctic ice sheets or loss of low-latitude coral reefs, and so the question becomes one about what the implications of a changed climate are at a more granular level. The aspect of interest in this thesis is the disruption to hydrological cycles. Through the Clausius–Clapeyron relation, an increase in air temperature results in an increase in its water holding capacity at a rate of approximately 7 % for every 1 °C rise in temperature. Consequently, the higher water vapor content in the atmosphere leads to an increase in precipitation (Boer, 1993; Pall et al., 2007). Furthermore, sea level rise has been occurring at a rate of approximately 3.6mm per year and is predicted to carry on doing so under a warming climate (Church et al., 2013; Haigh et al., 2017), due to a combination of thermal expansion of the oceans' water and an increase in the mass of water, through loss of ice from the cryosphere.

Putting these phenomena into context of the linked upstream and downstream impacts and what we have is an overall increase on the availability and flow of water through the Earth's systems. Over recent decades, there has been an increase in losses, both human and economic, to storms and flooding and although much of this is attributable to an increased human presence and capital exposure in areas of heightened risk (Changnon et al., 2000; Doocy et al., 2013; Few, 2003; Tanoue et al., 2016) it doesn't preclude future changes in risk. This risk may be driven, partly or primarily, by climate or in the short-term from the time lag effect between the release of GHG emissions and their associated impact (Ricke and Caldeira, 2014). Furthermore, deforestation and biodiversity losses, both of which continue to have a negative impact and show no sign of a significant slowdown (Butchart et al., 2010; Runyan

and D'Odorico, 2016), have been shown to exacerbate flooding (Gentry and Lopez-Parodi, 1980; Poff, 2002). If these trends continue, to say nothing of the feedback loops between deforestation, biodiversity loss, and climate change, then the risk profile may still change, albeit in a different capacity.

General consensus, according to the Intergovernmental Panel for Climate Change (IPCC), is that flood risk and drought risk will increase due to climate change (Field, C.B., et al., 2012), even if attribution of flooding impacts to climate at this stage is complex patterns are emerging (Blöschl et al., 2017). To add further complexity, the problem of future flood risk is clearly a non-stationary one, if it was ever stationary in the past (Wilby and Quinn, 2013); extrapolating from current trends is fraught with its own risk and future behaviour of the system should not be assumed based on statistical measures from the relatively short record of available observations (Hall et al., 2014). Therefore, one must have an idea of what the future climate, and perhaps other influencing factors, will be but, if one wished to predict said future climate change, then it should be noted that the Earth is best thought of as a highly chaotic system with significant potential for divergence from any anticipated series of events. A more nuanced approach is essential, especially when faced with the whims of human society.

1.2 Projecting Forward

The main tools at hand are General Circulation Models (GCMs), numerical simulations that represent the physical processes in the atmosphere, oceans, and cryosphere and those across the land's surface, being those that can affect the climate. Such models can, therefore, enable the exploration of future climate change scenarios when forced with different greenhouse gas and aerosol emission patterns. Five narratives (O'Neill et al., 2017; Riahi et al., 2017) have been developed for characterising challenges faced by society in adapting to or mitigating climate change. At a high level, these scenarios are:

SSP1 - Sustainability, a narrative focused on sustainability and inclusive global development, resulting in low challenges to climate change mitigation and adaptation.

SSP2 - Middle of the road, which essentially sees the world maintain a pathway not too dissimilar from historical trends, resulting in medium challenges to climate change mitigation and adaptation.

SSP3 - Regional rivalry, where nationalism and concerns over resources and security prevent broader cooperation and prioritisation of environmental concerns, resulting in high challenges to climate change mitigation and adaptation.

SSP4 - Inequality, of economic power, technology, and policy, that leads to improvements in clean technology but a lack of protection for some parts of society, resulting in high challenges to climate change adaptation but low challenges to mitigation.

SSP5 - Fossil fuelled development, a free market based approach, characterised by exploitation of natural resources to power global economic growth, resulting in high challenges to climate change mitigation but low challenges to adaptation.

The construction of the SSPs involves considerable detail on key socioeconomic factors, such as inequality and development indices, whilst still being versatile and enabling comparison, particularly with regards to the social aspects of climate change. When these trajectories of future global development are combined with those for radiative forcing (van Vuuren et al., 2014), the resulting emissions scenarios are plausible and justifiable futures to force the constituent GCMs of the sixth phase of the Coupled Model Intercomparison Project (CMIP6) (Eyring et al., 2016).

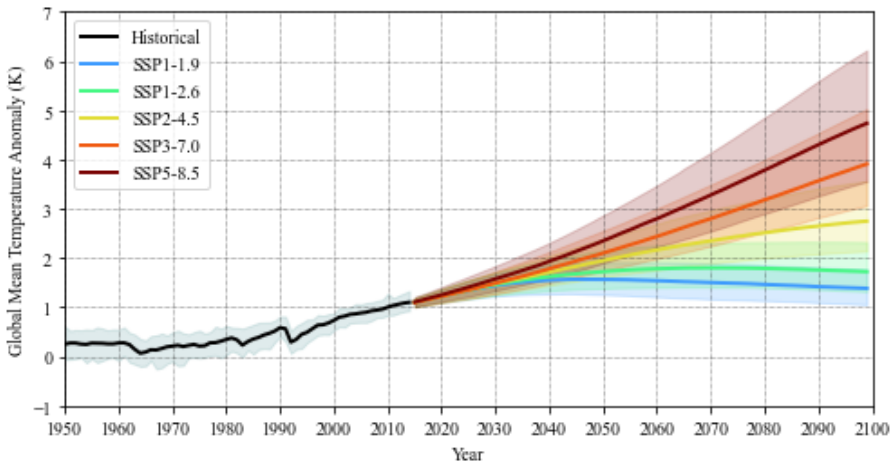


Fig. 1.2 Mean annual global surface temperature anomaly for each of the Shared Socioeconomic Pathway Scenarios

For these future scenarios, the predicted change in mean annual global surface temperature anomaly, where anomaly is the difference in temperature between the point of interest and a datum, using the multi-model ensemble mean, is shown in Figure 1.2 (V. Masson-Delmotte et al., 2021); the datum in this case is the pre-industrial average temperature. The more pessimistic scenarios lead, accordingly, to higher global surface temperature anomaly, with

the mean for SSP5-8.5 being approximately 5 °C, whilst the more optimistic scenarios, particularly SSP1-1.9, see project the temperature anomaly as stabilising at around 1.5 °C.

Those more optimistic scenarios, specifically SSP1-1.9 and SSP1-2.6 would be achieved via a rapid reduction in the use of fossil fuels and and increase in carbon sequestration. However, given the limited success of geopolitical developments, including that of the latest Conference of the Parties (COP) in Glasgow, 2020, the optimistic scenarios could be out of reach and the moderate to high emissions scenarios are perhaps more likely to better represent the actual future emissions. Under those scenarios it is more likely that the world will warm as little as 2°C and as much as 5°C or 6°C. The subsequent impact and change on present day climate is, therefore, higher and necessitates understanding both at a more granular level.

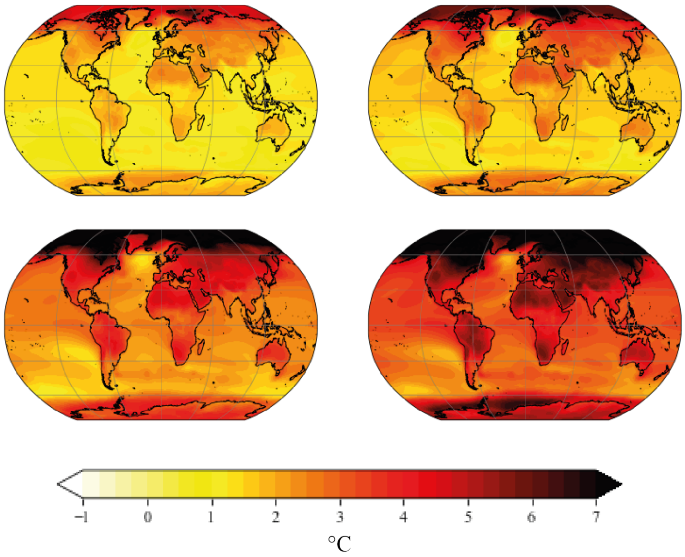


Fig. 1.3 Mean annual surface temperature anomaly for 1.5°C (top left), 2°C (top right), 3°C (bottom left), & 4°C (bottom right) warming, compared to pre-industrial levels, derived from CMIP6 multi-model ensembles (V. Masson-Delmotte et al., 2021)

Looking in detail at the temperature and precipitation patterns for 1.5°C, 2°C, 3°C, and 4°C warming and there is projected to be a general increase in temperature anomaly across the world, though significantly more so for the polar regions. For precipitation, on the other hand, there are both regions of increasing and decreasing relative intensity, with some regions potentially experiencing increases in intensity of up to 50% or more. It is easy, then, to understand the belief why greater instances of flooding may occur in some regions, especially

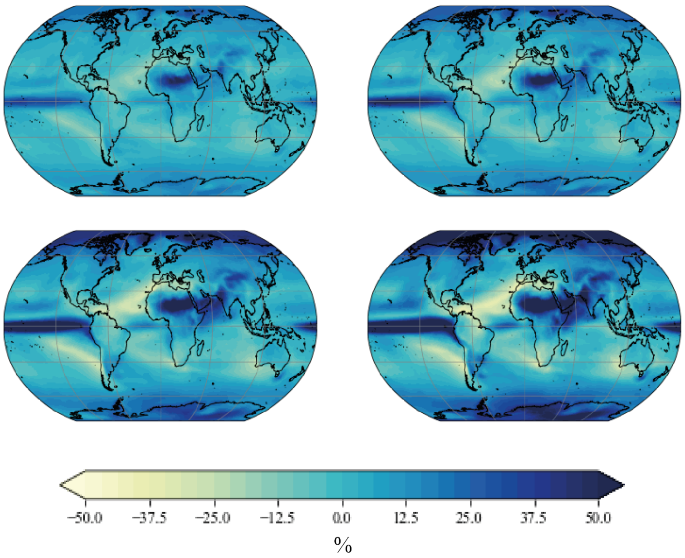


Fig. 1.4 Mean annual change in relative precipitation for 1.5°C (top left), 2°C (top right), 3°C (bottom left), & 4°C (bottom right) warming, compared to pre-industrial levels, derived from CMIP6 multi-model ensembles (V. Masson-Delmotte et al., 2021)

when some of the major flood events of past years, including repeated flooding in the UK, the 2021 Western European floods, and 2018 floods in Kerala, India, were driven by high intensity rainfall (Cornwall, 2021; Hunt and Menon, 2020; Spencer et al., 2018); further increases in the likelihood and severity of wet days could drive more frequent or severe flood events. Interestingly, whilst recent climate change has potentially weakened low-pressure monsoon systems, like those affecting the Asian subcontinent, this will be more than offset by moistening of the tropical troposphere under the worst warming scenarios.

Consequently, high level analysis of the impact of climate change suggests that, as anthropogenic emissions continue to rise, then the impact of precipitation intensity and the subsequent impact of flooding will also rise (Dottori et al., 2018; Prein et al., 2017). Thus, being able to quantify changes in hydrological behaviour across most geographies is critical, at a local level, subject to different emissions scenarios, such that a picture of evolving flood risk can be obtained and planned for.

1.3 A Paucity of Data

Imagine a context in which the risk from a disrupted global hydrological cycle is of interest for a specific region. Presumably significant study has already occurred and there is sufficient

data with which to use as inputs to or calibrate the parameters of either some mechanistic or empirical model. But what if that wasn't the case?

Hydrological data gathering is more focused on a subset of all countries, with those in South America, Africa, and parts of Asia having significantly lower gauging in place when compared to Western nations; this paradigm extends from climatic data gathering stations, such as for precipitation, to more involved hydrological process data gathering, such as river gauging (Kidd et al., 2017; Krabbenhoft et al., 2022). Overlapping this with information on communities that are more susceptible, either through increased probability or severity of climatic events or though lower resilience or poorer health outcomes (Doocy et al., 2013; Tanoue et al., 2016), and the consequences exacerbate already bleaker outcomes.

To compound the matter further, beyond what data is available to how that data is actually viewed, the bias within rain gauge datasets has been found to lead to further bias in the variability associated with subsequent estimates for extreme precipitation (Sun et al., 2018). Assuming this bias exists with similar hydrologically focused data and without sufficient data to ensure fair representation, when it comes to making data adjustments or interpretations, particularly for threshold exceedance or anomalous events, the practitioner essentially tends toward weighting the performance of their methods more heavily in favour of data they are presented with rather than speculating. In other words, bias through underrepresentation in key datasets exacerbates downstream biases the limited data picture available for some regions becomes even more so.

An obvious fix is to procure more data but data gathering is, however, an expensive, or sometimes unfeasible, activity. Peppering the underrepresented regions with streamflow and rain gauges is both unlikely and would also do little to solve the lack of sufficient historical data. Even if stationarity with respect to climate and human-environmental interactions was assumed, contrary to our point before, then making rational projections within that context would be impossible due to limited records (Stedinger and Griffis, 2011).

1.4 Rationale

If certain geographies have ample data whilst others have a dearth of it, then a suitable methodology for extrapolation is required and, if performing this extrapolation for better spatial coverage, why stop there? Why not attempt to employ a method that is suitable for handling temporal non-stationarity? Assuming that the patterns and permutations of and relationships between independent variables for those regions considered to have adequate data can be considered broad enough to represent permutations of those without and that the records available capture enough heavy flooding and prolonged drought events, then pattern

recognition models may be able to generalise better than simply extending trends (Bishop, 2016; Murphy, 2012).

Having established that data paucity is likely an issue with regards to predicting future climate impact, the first argument for this thesis is that, by levering complex patterns occurring within the data, machine learning algorithms can offer an empirical route to generalising hydrological models that offer robust predictions for streamflow and subsequent flood risk. A more extensive introduction to machine learning is offered in Chapter 4. Our approach will be built up from first principles whilst examining the relevance of all features used an inputs to the model, such that the machine learning framework is as lean as possible and prioritises the use of data that can be collected externally from the catchment. This maximises the capability of the framework to generalise, extrapolating patterns from gauged catchments to those that are not. An approach from first principles also offers an opportunity to evaluate a machine learning model's capability to capture and internalise intermediary processes and behaviours, such as the antecedent conditions, without needing to explicitly input or represent them.

Further elaborating on the nature of data paucity, some of the data types used here are expected to be spatially and temporally complete, such as the meteorological data or topographical data, the potential existence of spatiotemporal biases within those datasets notwithstanding. Expected sparsity that we will redress lies within spatial hydrological dimensions, to wit: the lack of gauging in some geographical locations. The set of catchments considered temporally complete, those having a significantly long record of hydrological data gathered, will constitute the machine learning training sets, whilst those entirely ungauged or significantly incomplete will form the test sets. In order to test this hypothesis, we will take a series of catchments from the UK, where hydrological measuring and gauging is extensive, and treat a subset as though it were ungauged, using its streamflow data only for comparison against predictions. Machine learning is highly suitable in this context, both as a framework for handling sparse datasets and as a modelling application that can eliminate the need for prescribing complex parameterisations, thereby enabling us to remain relatively agnostic about the interactions between the forcing variables.

A subordinate hypothesis to the above is that the domain knowledge required to train the machine learning algorithm is of more import that the architecture; however, we still would expect the structure of some machine learning architectures to be better suited to modelling this problem than others, such as sequence, or times series, models. Additionally, emphasis on the relative nature of our being agnostic to the interactions between the forcing variables is another aspect of exploration within this thesis; for example, if we know that increasing precipitation results in higher streamflow events, then this might be a condition that should

be enforced to ensure out of training set performance. Ergo, a range of models are explored that range from parametric models invariant to permutations within a training set example, such as multi-layer perceptrons, through to Gaussian Process methods that allow for the expression of prior knowledge or beliefs. Our aim is to determine the impact of model choice on performance; by extension, we also consider whether or not custom architectures, based on adaptations to existing architectures, tailored towards specific problems lead to superior performance over the more ubiquitous.

The second aim is for the development of machine learning algorithm that can be used to identify trends for extreme meteorological events, specifically the subset having significant influence on hydrological events. Due to data availability and veracity, discussed further in Chapter 3, the focus will be on widely reported synoptic scale events; this sites the application to an area with a high level of data collection, the North West Pacific Basin. Whilst this might seem incongruous with the preceding chapters, our aim is to underline how patterns in circulation data can be recognised by machine learning and whether or not resulting impacts can be predicted from this high level data; having access to data is key to showing this.

More formally, our hypothesis here is that a computer vision type machine learning model can identify circulation patterns from meteorological and climate model data arranged in at least three dimensional tensors in order to make probabilistic predictions for the evolution of extreme phenomena and their immediate impact. Given that our focus in on hydrology impact, the expectation is that this can then act as an input to the hydrological models explored under our first main objective. In other words, we might be able to capture the significant impulse of extreme meteorological events, such as synoptic scale storms, from a coarse representation of general circulation, take it an interim output, and then make a prediction as to their subsequent flood risk.

Reframing these prior objectives against the wider issue of climate change and projections thereof, if we are able to pre-train the models on historic observations and make an appropriate transformation from climate model data, then the resulting system ought to be one capable of generating granular statistical distributions of hydrological impact under the different emissions scenarios. The expectation for this overarching framework is that it can enable climate change impact modelling and robust predictions to counter data paucity whilst being computationally efficient.

Suppose that the objectives of the first two aims are suitably met and the platform is indeed considered robust, then it would be more difficult to directly assess the merits of the model in terms of its accuracy in generating climate model impacts, given that the data by which to make the comparison doesn't exist. However, under our assumption. the answer as to the suitability of the models is already borne out and their performance when forced by

meteorological data meets expectations. Instead, we will comment on standard transforms of climate model data in terms of their suitability for forcing a climate impact model. Ergo, our third and final hypothesis is that some common methods, particularly linear transforms, of climate to meteorological data are not suitable; a new framework might yield better overall performance and be better paired with the other elements of this thesis and our aim is to highlight a new direction.

1.5 Outline and Contribution of Thesis

The main contribution of this thesis is to develop machine learning frameworks for generating future streamflow predictions, including a set of models able to perform inference based on one set of locations and generalising to another and the introduction of hybrid probabilistic models for time series phenomena.

In Chapter 4, we introduce a general linear model alongside a basic artificial neural network, both using the assumption that it might be possible to express streamflow as a function of rainfall. This is expanded upon through more involved feature engineering, to generate more accurate predictions, whilst also evaluating the use of hydrological state, in the form of soil moisture and storage; this internal state can be internalised with a machine learning model and even replaced with a set of climatic proxy variables.

Following on from the feature engineering, we open Chapter 5 with a comparison of different model types and their application to the problem, including neural network types, Gaussian Process types, and hybrids. Whilst extreme behaviour can be accounted for in the kernel design of a Gaussian Process, it is harder to enforce this in neural networks, so an approach for better fitting to extremes is also included here.

The problem is extended from a single catchment to multiple catchments in Chapter 6, where we use a machine learning model to learn from one set of catchments in order to predict for another set. The initial setup involved using physical catchment descriptors, such as area or land use; model performance was improved by including additional variables to approximate human impact on hydrology.

Chapter 7 describes two extensions of the Neural Process model to sequences, the Recurrent Neural Process, based on a recurrent neural network, and the Temporal Process, incorporating a temporal convolutional neural network for causal based inference. We use these models in the single catchment setting to highlight their potential against existing models used in Chapter 5.

Chapter 8 applies machine learning to the detection and quantification of extreme weather events that may cause extreme precipitation and subsequent flooding, specifically storms.

Using typhoon data, we use convolutional networks to identify and track storm activity and provide a prediction for precipitative impact.

Finally, in Chapter 9 we apply the most accurate models developed to future climate data to generate a series of potential impact studies of a changing climate under the different SSPs, highlighting the change in streamflow distributions with respect to emissions and the suitability of driving machine learning models with GCM data.

Chapters 2, 3, & 10 are ancillary chapters, through which we provide: a review of prior art in hydrology and related fields in Chapter 2; a conceptual and theoretical basis along with input and output data sources and extraction methodology in Chapter 3; and our conclusions, summary of findings, and final remarks in Chapter 10.

Chapter 2

Prior Art

2.1 Hydrology

All models that describe natural systems can be broadly categorised as empirical, being those that are derived from an observed relationship between the input and the output, or mechanistic, being those that are based on the understanding and mathematical expression of physical laws or processes. Hydrological models are no exception, although a further subdivision is often used to refer to models that exist somewhere between fully empirical and mechanistic models, conceptual models that use a simplified mathematical representation of the system (Devia et al., 2015; Jaiswal et al., 2020).

A range of mechanistic models have been developed, leveraging the advent of more powerful computers to shift away from simplistic two dimensional models (Beven, 2001), that solve the differential equations key to hydrology for a catchment; prominent examples include the Système Hydrologique Européen (SHE) (Abbott et al., 1986a,b), its derivation, MIKE SHE (Singh, 1995), the Soil and Water Assessment Tool (SWAT) (Arnold et al., 1998; Srinivasan et al., 1998), or the comprehensive and more recent HydroGeoSphere (HGS) (Brunner and Simmons, 2012). These models enable a broad range of internal processes to be simulated, such as groundwater recharge, components of the nitrogen cycle, or evapotranspiration with a high degree of skill (Cornelissen et al., 2016; Golmohammadi et al., 2014; Jaiswal et al., 2020) which is a key strength of this type of modelling. This high skill level is unsurprising but through the further analysis of such models, there is a requirement for user skill and appropriate data to calibrate the models appropriately; this might serve to hinder the application of these models in a global context. Furthermore, it has been noted that fundamental issues with regards to process representation and model structure often go unaddressed (Refsgaard et al., 2010).

As far as empirical methods are concerned, a gold standard, certainly within industry, is the Flood Estimation Handbook (Calver et al., 2009; Faulkner et al., 2012; Institute of Hydrology, 2008; Samuels et al., 2008), a set of statistical methods that enable the generation of design floods and peak flow estimation through assimilation of observational data and hydrological similarity between catchments. Whilst the error metrics for these methods demonstrate a high degree of skill, our concern is that the methods have a frequentist leaning that might not hold up well in the face of a nonstationary climate and the tighter demands made of predictions in response (Peel and Blöschl, 2011). Although it might be easy to think of empirical models, and in particular machine learning models, as being data intensive compared to conceptual and mechanistic models, both conceptual and mechanistic models require a large amount of hydromorphological data for calibration or boundary conditions (Devia et al., 2015; Jaiswal et al., 2020). In order to assess a model the output also has to be measured, even if not directly required to parameterise the model; therefore, a properly designed machine learning modelling approach can be applied without necessarily entailing significantly more work than conceptual or mechanistic approaches. Furthermore, if the physical and mathematical relationships are improperly parameterised, and we note that the perfect mathematical representation of complex physical systems is likely impossible (Lorenz, 1969), then machine learning models perhaps offer an opportunity to learn more about a system or create predictions when internal information is unavailable. Consider the Aral Sea, a prime example of a hydrological system that has been heavily influenced by human activity to the point of ecological crisis (Loodin, 2020; Micklin, 2007, 2010); yet, despite it being a focal point of study due to its size, importance, and anthropogenic forcing, data on water use is scarce across much of its catchment, such as that from Afghanistan, which encompasses 12% of the Aral Sea Basin, in particular being incomplete and unreliable. If we assume that patterns of water use, including agricultural usage, are driven by climate then some of these behaviours might be internalised within a machine learning model.

One of the major problems, as was highlighted in Chapter 1 with regards to paucity or sparsity of data (Kidd et al., 2017; Krabbenhoft et al., 2022) and above, is that flood problems can and indeed do occur in the absence of data (Robson and Reed, 2008), where no gauging is conducted and no data exists with which to calibrate certain model classes. Compounding matters is that we are living in a changing climate, one where we now have to make extrapolations or predictions across both space and time (Peel and Blöschl, 2011). It is in these gaps that the potential of machine learning might be best applied, in extrapolating patterns or relationships (Bishop, 2016).

Artificial Neural Networks (ANNs) have been used in hydrology for a significant period of time, since their potential was enabled through the development of backpropagation, with

extensive meta-analysis on the application of machine learning in the field being conducted (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000; Govindaraju et al., 2000). In spite of the this, uptake and credibility has been hampered somewhat by a lack of consistency in data frameworks and transparency (Abrahart et al., 2012) but the issue of a lack of transparency in terms of model mechanics does not seem too distant from the process representation issues presented for mechanistic models (Refsgaard et al., 2010). More recent application of ANNs has resulted in predictive performance that approaches perfect agreement between predictions and observations, depending on the complexity of the model employed, both in terms of the inputs and the architecture (Aichouri et al., 2015; Ali and Shahbaz, 2020). Thus ANNs must be internalising the catchment physics to a certain degree.

The nature of the problem, in the clear temporal element across input variables, has led to sequential or time series adaptations of neural networks being applied with a high degree of success, either in the Recurrent Neural Network format or the Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) derivations (Ayzel and Heistermann, 2021; Gao et al., 2020; Nagesh Kumar et al., 2004). If we consider the wider field, then more recent models, such as a temporal adaptation of the Convolutional Neural Network, are being applied to environmental data science problems, such as in the successful prediction of the El Niño-Southern Oscillation (ENSO) (Yan et al., 2020), given the temporal nature of many of the problems within the field.

Beyond the context of modelling streamflow at a single outflow, even though the use of machine learning for generalising hydrological models has been questioned (Devia et al., 2015), machine learning algorithms have been applied to predict streamflow at different gauging sites within the same basin (Besaw et al., 2010) and even extended beyond the confines of a single location to generalise across catchments in the United States with mean results meeting a benchmark for industrial relevance (Kratzert et al., 2019), even if not on par with the best single catchment or mechanistic models. A direct comparison has been made between the Flood Estimation Handbook and ANNs for the purpose of predicting peak flows using what would be termed a small dataset by contemporary standards, with ANNs having an edge depending on assessment metric, size of catchment, and structure of the dataset (Dawson et al., 2006). These approaches could be further refined in terms of developing a framework that generalises more readily, particularly in the face of human intervention within a hydrological system and a lack of data availability. Furthermore, there is room to expand the scope in terms of the machine learning framework adopted. Thus, there is clear potential for machine learning algorithms in this space to be further refined and deliver generalising models with the right approach.

2.2 Storms

Of natural disasters, storms are visibly impactful, can be highly damaging, and, crucially, pose significant flood risk, making them the second most damaging natural disaster (World Bank and United Nations, 2010). Hurricane Katrina, making landfall at the South East coast of the United States, and Typhoon Haiyan, hitting the Philippines, resulted in 1,836 and over 6,300 fatalities, respectively, and a combined US$140 billion in damages, inundating vast swathes of land (Lagmay et al., 2015; Schiermeier, 2013). Haiyan, in particular, caused devastating storm surges, fuelled by heavy rainfall and high wind speeds.

The impact of tropical and extra-tropical storms in the future is uncertain but, thus far, the human impact depends on the intrinsic properties of the storm along with social and political factors, not least of which is the exposure of human infrastructure and development along shorelines (Changnon et al., 2000; Peduzzi et al., 2012; Woodruff et al., 2013). If, then, coastal development continues to put humans in harm's way and is further exacerbated by an increase in sea-level rise, and that occurs without there being a change in trends in storm activity, then overall risk of loss due to storms could likely worsen and storm prediction over long time periods is a necessary activity in driving adaptation and mitigation strategies.

Over the past few decades, the short range forecast of cyclone tracks, up to 120 hours, has been improving significantly, with the error decreasing by a factor of more than 4 since the 1960's (Cangialosi et al., 2020; Landsea and Cangialosi, 2018), however, there is an assumed limit to the advances in predictive capability, due imperfect information about the system, particularly subgrid scale processes, that results in increasing divergence over time (Lorenz, 1969). These models solve fluid dynamics equations that are now at a fine grid scale with physical processes being well parameterised to deliver track errors, taken at the centre of a cyclone, of approximately 50km for the first 24 hours (Chen et al., 2019). The location of a storm is well captured and predictions of intensity, in terms of wind speed, is improving (Rogers et al., 2013), although predictions of rainfall or subsequent storm surges (Son et al., 2022) are areas needing further attention. Ultimately, our impression is that predictions and projections for storms in the short term are robust and modelling capability here is strong, albeit computationally expensive. We note that there could be the potential for machine learning to provide probabilistic forecasts to supplement these physically driven models, such as was done with sea ice predictions using Convolutional Neural Networks, specifically U-Nets (Andersson et al., 2021).

Obviously, being able to provide exact predictions for storm forecasts over a longer time frame is impossible; even predicting the location of where a storm might occur is complex, given that a conclusive theory of cyclogenesis, being the formation of cyclones, and the driving physical processes is lacking (Horn et al., 2014; Lee et al., 2020; Ruppert

et al., 2020; Walsh et al., 2016). There are indices for the potential of cyclogenesis, most of which have relatively robust performance in terms of spatial patterns and seasonal variability but are misaligned with directly detected cyclones (Cavicchia et al., 2022); we might hope that the empirical nature of machine learning models, along with the potential to extract feature weighting or maps from parametric machine learning models, could be of use in both predicting and locating cyclogenesis (Qian et al., 2022) and unlock further insight in this area. Shortfall in a perfect understanding of cyclogenesis aside, cyclones are a behaviour represented within climate models. Future projection, therefore, becomes a matter of examining statistical metrics of potential changes in genesis potential, cyclone frequency, and cyclone intensity over long time scales using atmospheric models, a practice that has a long history (Wehner et al., 2017). The ability of models to represent cyclones that reflects observations is dependent on physical process representation along with resolution (Camargo, 2013; Roberts et al., 2020a), though aspects of the modelling setup, such as numerical scheme (Reed et al., 2015), also contribute to the capability of models to accurately represent cyclonic activity. Consequently, as models have become more refined, both in terms of model physics and resolution, the error between models and observations has closed when compared with previous iterations although some regional biases still persist (Camargo, 2013; Emanuel, 2021; Priestley et al., 2020; Roberts et al., 2020b).

Limited consensus has formed around the futures affected by anthropogenic climate change with regards to cyclone intensity and frequency, with suggestions of both increases and decreases across different studies (Emanuel, 2021; Roberts et al., 2020b; Wehner et al., 2010; Yoshida et al., 2017). Some of the discrepancy has been attributed to the use of tracking algorithm (Horn et al., 2014; Roberts et al., 2020b), for which multiple frameworks exist to detect tropical storm tracks in GCMs, and other high resolution atmospheric models, using the spatial distribution of atmospheric temperature, windspeed, and moisture (Bosler et al., 2016; Knutson et al., 2008; Vitart et al., 1997; Vitart and Stockdale, 2001; Zhao et al., 2009). Again, we question whether or not machine learning can be introduced here to automatically detect, track, and make inferences about cyclones from relatively high resolution data, perhaps even improving predictions of precipitative impact and circumventing the use of regional climate models that have previously been used to predict storm related precipitation (Liu et al., 2019b; Wright et al., 2015).

If we shrink the scale even further, extreme phenomena, such as tornadoes, convective storms, or hail, can occur that are typically below the resolution of extratropical and tropical cyclones and are difficult to measure. In real time, their prediction often relies on observational reports from a range of sources of variable veracity with density being skewed towards areas of higher population centres (Allen, 2018; Edwards et al., 2018; Groenemeijer et al.,

2017; Verbout et al., 2006). However, the compilation and assimilation of these datasets along with processes to counter prevalent overestimation can enable statistical relationships between extreme phenomena and the spatiotemporal distributions of climatic variables to be established and reporting accuracy has significantly improved over time (Ostby, 1999; Taszarek et al., 2020); indeed, it is possible to identify the conditions that lend themselves towards some of these phenomena from meteorological reanalysis datasets (Rädler et al., 2018; Westermayer et al., 2017). Considering that the genesis and evolution of these events might be complex and difficult to parameterise, there exists the potential for machine learning methods here to help drive learning of these relationships and perhaps make probabilistic projections of occurrence at particular scales of model.

2.3 Climate Impact Modelling

In the context of climate change, adaptation and mitigation strategies are required around changing water resource and temperature distributions, whether of large infrastructure systems or human health in the face of changes to extreme phenomena frequency and intensity or disease distribution (Allard, 2021; Laukkonen et al., 2009; Tabachnick, 2010). Understanding the nature of suitable adaptation and mitigation strategies is complex and requires robust and granular climatological information (V. Masson-Delmotte et al., 2021); the kind of granular information required to act as a forcing for such downstream impacts is below grid scale, necessitating the development of localised models as the intermediate step between high resolution climate models and these downstream models. In other words, we wish to know how certain aspects of the geophysical system, not captured in GCMs, respond to climate change such that we can take their response and propagate it into models that rely on them to predict a subsequent impact. Furthermore, commonly used impact frameworks by policymakers, such as integrated assessment models, are too simplistic to present a robust picture surrounding climate impact and the risks that it entails (Ackerman et al., 2009; Weyant, 2017); even at the engineering level in hydrology, by using frequency analysis methods based on historic observations, the simulated impact is unreliable, treating climate as stationary when it is not (Brown, 2010; Forsee and Ahmad, 2011; Milly et al., 2008). This is further exacerbated when we realise that the impact to infrastructure does not have a linear relationship with the intensity of an event (Allard, 2021; Auld, 2008; Emanuel, 2021) and design levels or factors of safety must be weighed against probability distributions of risk. Essentially, those tasked with mitigating climate risks are being underserved by the information they rely on. More powerful modelling approaches that leverage projections of climate whilst remaining

accessible are essential for understanding the impact of climate change and for propagating risks from the geophysical systems to our human ones.

As we have alluded to before, Climate models are based on the physical laws of conservation of energy, mass, and momentum, as well as thermodynamic and radiation laws but these models are still simplifications of a large and complex natural system. That there should be some error is to be expected when considering their relatively coarse scale and that the representation of all physical processes is impossible. Furthermore, these simplifications, or parameterisations, can affect other, non-close processes, being far in terms of either the spatial or temporal domain (Maraun et al., 2017; Tebaldi and Knutti, 2007; Wang et al., 2014). Thus, when we consider that these models are not designed to represent interdiurnal variability of meteorological variables on the same scale as short range weather forecasts, their output is not suitable for directly forcing impact models.

Downscaling and correcting GCM model output can either be done through a dynamical method, that is to say through a still mechanistic yet finer resolution Regional Climate Model (RCM), or through a statistical method (Maraun and Widmann, 2018; McGregor, 1997). Using a dynamical model, however, requires complex physical parameterisation and, therefore, may be at odds with a goal of creating usable data at all points within the domain, even if they are adept at simulating local climate; additionally, an RCM can suffer from similar bias and calibration issues to those of a GCM with post processing required on top of the already high computational cost and complex parameterisation (Maraun et al., 2010; Teutschbein and Seibert, 2012).

Within the field of statistical downscaling methods, commonly used techniques include delta change field, model output statistics, perfect prognosis, and weather generators (Maraun and Widmann, 2018; Tabari et al., 2021). Delta change field techniques and model output statistics are transforms of the GCM output using, as the name might suggest, statistical measures, such as the mean. Weather generators and perfect prognosis, on the other hand, involve the application of a statistical model; for weather generators, this is based on using a statistical distribution to create time series data and, for perfect prognosis, a statistical model that defines a relationship between large-scale and local-scale variables before being applied to the large-scale output of GCMs.

Of these techniques, model output statistics, particularly in the form of Quantile Mapping, have been applied to mechanistic hydrology as a form of bias correction, with corrective capability deemed reasonable when it comes to the statistical distributions of precipitation (Bian et al., 2021; Chen et al., 2013); where information is severely lacking, delta change methods can be employed (Buytaert et al., 2009). Both of these methods feature relative ease of application and computational simplicity but there is a risk that they fail to properly capture

the change in the parameterisation of the statistical distribution of meteorological variables and the interdiurnal variation that might characterise future climate scenarios, an issue that we will explore here. With regard to computational simplicity, however, we note that generative machine learning models are capable of producing physics informed time series (Kashinath et al., 2021; Lim and Zohren, 2021) and the relative computational efficiency of inference could lead to the development of times series data with which to force impact models.

Chapter 3

Theoretical & Data Foundations

This chapter is provided as a conceptual and theoretical grounding for the work carried out in this thesis whilst also providing a reference point for all of the data used throughout, in terms of its origin, veracity, and processing.

3.1 The Hydrological Cycle

Hydrology is essentially the science of water as it occurs in the environment, in terms of its distribution, movement, and related properties (Chadwick et al., 1998; Dooge, 1973; Raudkivi, 1979); as a result, a wide range of fields could come under the umbrella of hydrology, such as cryology or oceanography, although in this thesis we are concerned with the fluviatile subset of processes.

If we consider a section of land bounded by topography with surface flows feeding a single river and any tributaries, such that these outflows converge at a single point, then this is what we refer to as a catchment. A catchment may or may not feed directly into a sea or ocean but the overall cycle is characterised by the transfer of water primarily between atmosphere, oceans, and components of the catchment subsystem, which we have represented via a process diagram in Figure 3.1.

The catchment system, in terms of inflows, outflows, and internal movement of water has to obey conservation of mass and the factors that influence these processes, including but not necessarily limited to: the catchment area, soil type and depth, local geology, vegetative cover, land usage, lakes and reservoirs present within the catchment, the local drainage network both artificial and natural, and the topography, specifically stream and surface slopes (Chadwick et al., 1998; Dooge, 1973; Raudkivi, 1979; Yilmaz et al., 2008). The sum total of water that enters and leaves the catchment through climatic processes, that is discharged by a river, and that is stored within a catchment must be equal to zero, as per Equation 3.1, where

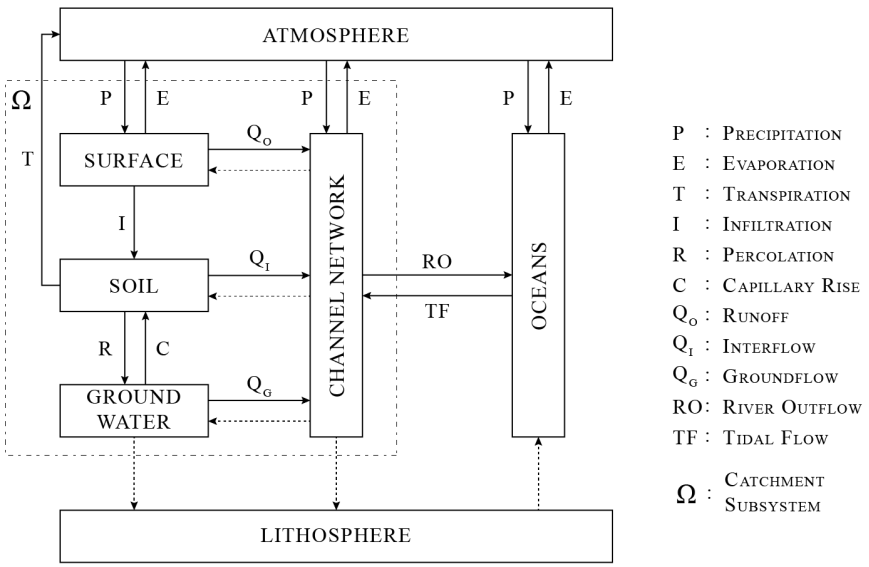


Fig. 3.1 Hydrological process diagram, representing capacitance as blocks and connecting processes, adapted from (Dooge, 1973) and (Raudkivi, 1979).

the cumulative precipitation over a time period with length T in the catchment is given by $\int_{t=0}^{T} \Psi_t dt$, the water leaving the catchment through evapotranspiration is $\int_{t=0}^{T} E_t dt$, the target streamflow exiting through the river is $\therefore \int_{t=0}^{T} Y_t$, and the change in water storage is ΔH.

$$0 = \int_{t=0}^{T} \Psi_t dt - \int_{t=0}^{T} Y_t dt - \int_{t=0}^{T} E_t dt - \Delta H \tag{3.1}$$

$$\therefore \int_{t=0}^{T} Y_t dt = \int_{t=0}^{T} \Psi_t dt - \int_{t=0}^{T} E_t dt - \Delta H$$

When it comes to a flood, the temporal component of these variables becomes critical, with intense climatic conditions, specifically precipitation, potentially resulting in significant streamflow events. The manner in which this manifests is through a significant impulse in streamflow, traditionally referred to as peak flow above base flow. The other components of the equation can either assuage or compound the issue, such as the internal storage mechanisms of a catchment already being saturated.

However, the full characterisation of a hydrological system is obviously complex, thus the representation of a hydrological system on both a conceptual and practical level must be addressed; in this, there exists the potential for compressing or simplifying the catchment to decrease both the information gathering burden and the system parameterisation. A broad distinction that we can make is to categorise models as either being 'lumped', where all of the input variables are aggregated spatially, or 'distributed', where all of the input variables are distributed spatially across the domain.

Distributed models enable the evaluation or representation of the spatial distribution of hydrological processes; the result being that, when correctly calibrated and presented with sufficient input data, they can offer insight into the interior flows and allow for the evaluation of flood behaviour at interior ungauged locations, i.e. upstream of a catchment's gauging points. However, in a mechanistic context, they require extensive calibration and, more generally, have not been shown to necessarily offer improved performance over lumped models whilst being difficult to parameterise (Reed et al., 2004; Smith et al., 2004; Yilmaz et al., 2008).

An obvious issue that arises is the obtainment of appropriate input data at the required spatial resolution. Notably, the higher resolution of models that comprise CMIP6 is up to 0.25°or 25km (Haarsma et al., 2016); therefore, if, for example, a catchment were to be fully contained within relatively few grid squares, or even just one as might be the case for smaller catchments, then a downscaling operation would be required to increase the spatial resolution to be compatible with a distributed model for that catchment. The same applies to other spatial catchment data. As such, given the aims and direction of this work, the use of distributed models might not be appropriate. Thus, in the first instance, a lumped modelling approach will be adopted, particularly if the end goal is to use CMIP data to force the hydrological model; our procedure for performing the compression of the system is detailed later within this chapter.

3.1.1 Extreme Phenomena

There are multiple types of extreme phenomena, most arising out of an energy imbalance between the fluid systems that comprise the atmosphere and oceans, with several having the potential for significant hydrological impact. Based on the theory outlined above, the extreme phenomena of relevance here are those that do have more hydrological impact and feature extreme precipitation, temperature, or wind speed. This would include phenomena such as tropical and extra tropical cyclones, convective storms, blizzards or snowstorms, wind storms or severe wind, tornadoes and whirlwinds, and droughts or heatwaves. A particularly broad goal would be classifying and identifying the impact of all types of extreme phenomena but,

naturally, our prior stated goal, as per Chapter 1, is to understand the capability of machine learning in a measurable and well defined subspace. Thus, the extreme phenomena we focus on is cyclonic storms. Our rationale here is twofold: first, the evident causal link to flooding, such as with the 2013/2014 winter flooding across the UK (Kendon and McCarthy, 2015; Muchan et al., 2015) or the extensive damage and loss caused by Hurricane Katrina and subsequent flooding in New Orleans 2005 (Jonkman et al., 2009); second, data availability and our data selection strategy, as detailed in the ensuing data sources section of this chapter.

In order to identify storm activity, or make any prediction about its movements or impact, one must first clarify the definition of a storm in this context and further elaborate on their nature. The storms we refer to here typically occur on the synoptic scale, being at horizontal resolutions of the order of 1000km, or at mesoscale alpha scale, ranging from 200 to 1000km and being close to the edge of synoptic scale, and are cyclonic, characterised by rotating air about a low pressure centre. A further division is made between those occurring within the tropics and outside of them, as the formation and structure differs between tropical and extra-tropical storms (Gray, 1979, 1998; Wash et al., 1990). Whereas extra-tropical storms are caused by interactions between bodies of air at different temperatures, tropical storms are formed when warm sea surface temperatures heat the air above in regions of low pressure. As winds converge into the region of low pressure and the heated air rises, drawing up moisture, the Earth's Coriolis effect generates spin in that rising air; as it rises, however, it also cools and the moisture condenses to form large, heavily saturated storm clouds. Essentially, tropical storms are heat engines that transfer thermal energy from the oceans into the atmosphere, with the resulting pressure differential further driving the storm, as long as that heat source remains. In other words, if the heat source is removed, such as the storm passing over land, then the storm can breakdown. We further characterise tropical storms as having spiralling bands of wind and rain about a low pressure centre, the eye of the storm. The precipitation and windspeed intensity typically increase radially towards the centre up to the eyewall, the area that borders the eye.

3.2 Data Sources

3.2.1 Meteorological Data

Meteorological Reanalysis

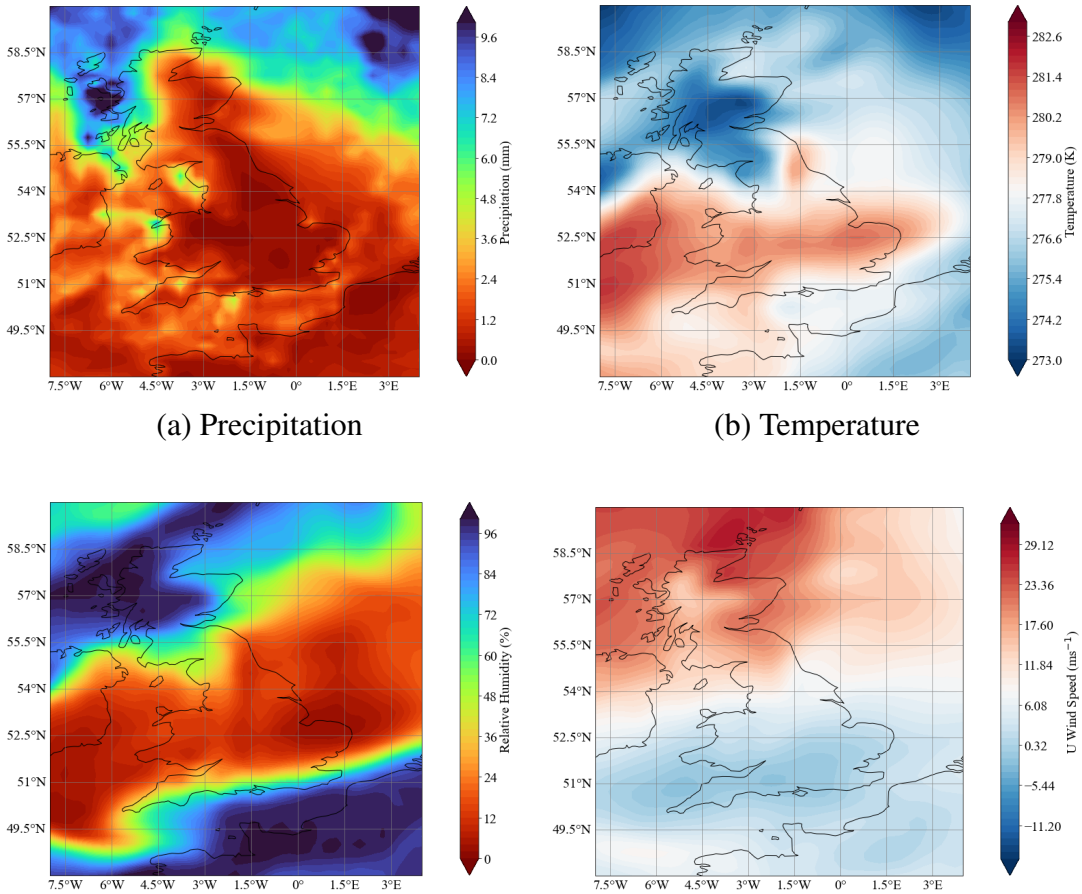
Accurate measurements or representations of meteorological data are critical to this thesis, with the primary variables of concern being precipitation, temperature, wind speed, and humidity. However, as was noted in Chapter 1, global coverage for measurement of all

meteorological variables is impossible and, in some cases, insufficient, particularly with regards to precipitation data (Kidd et al., 2017). The European Centre for Medium-Range Weather Forecasts (ECMWF) fifth generation of global climate reanalysis data, ERA5, is a reanalysis product that utilises observational data along with data assimilation and modelling capability to produce global estimates for meteorological data (Hersbach et al., 2018). ERA5 data is widely used due to its high accuracy and comparability with observations, with negligible difference on the output of hydrological studies (Hersbach et al., 2020; Nogueira, 2020; Tarek et al., 2020); the dataset even offers accurate representation of tropical cyclones (He et al., 2020), although there are some instances of regional bias (Jiao et al., 2021).

Ergo, ERA5, downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store, will form the primary meteorological data source for this thesis. The subset of data that will be used is the period from 1979 to 2019, bounded within longitude of -8° and 4° and latitude of 48° and 60°. The variables extracted from the data product are precipitation, taken at surface level, and temperature, relative humidity, and the components of wind speed parallel to the earth's surface, all of which have been taken at a height just above surface at 1000hPa. Example contours for an arbitrary time step are illustrated in Figure 3.2.

The pressure level variables could be extracted at a range of different pressure levels but, in order to limit the amount of data gathered, we note that the spatial distribution patterns are similar across the different pressure levels; with that in mind, we would expect a machine learning algorithm to be capable of learning comparatively well on a subset of all available pressure levels. Furthermore, there are feedback mechanisms between the different pressure levels, meaning that influences between pressure levels can be captured either forward or backward in time, and the variables at pressure levels closest to the surface will have a greater impact on relevant processes, such as evapotranspiration.

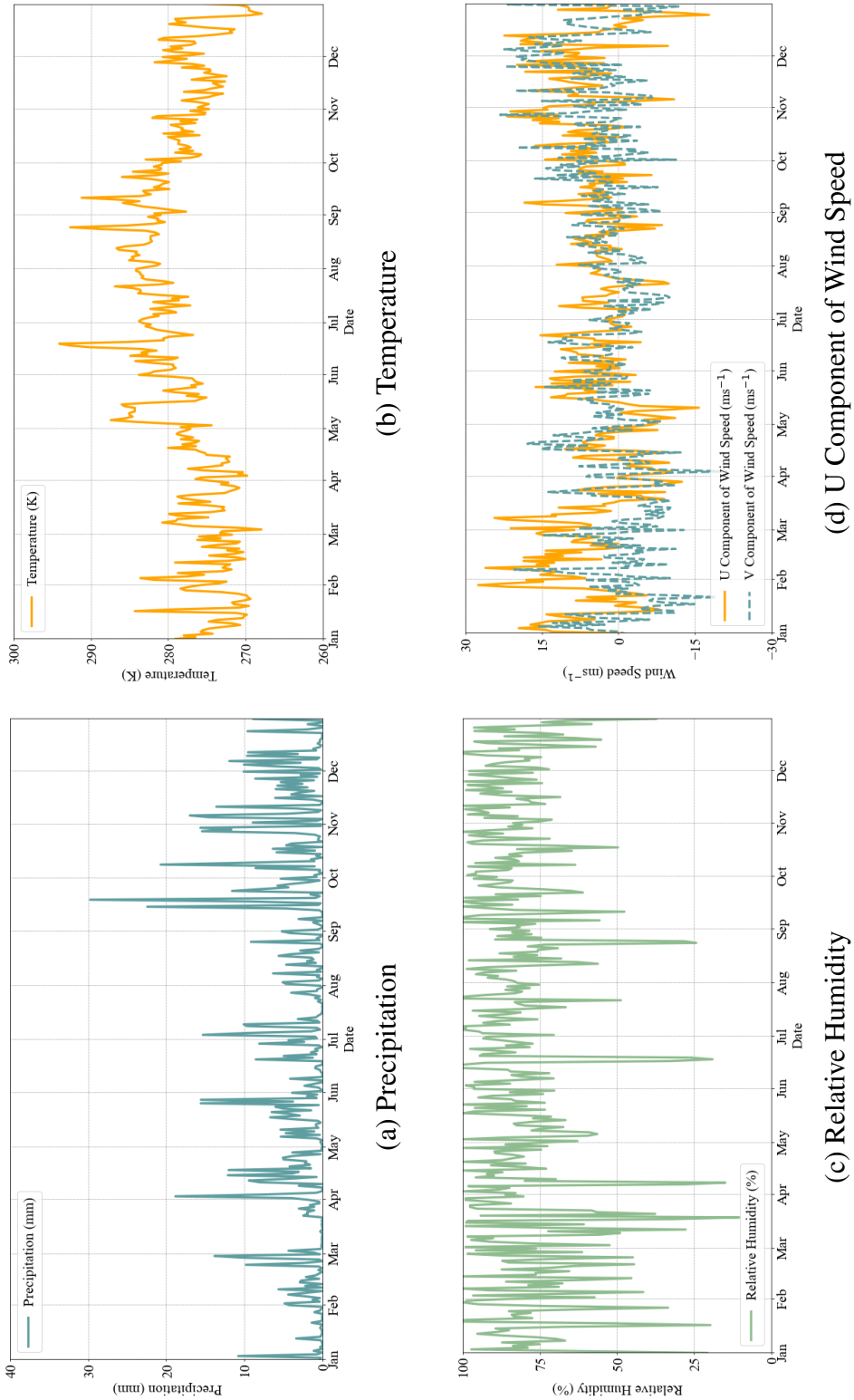
In order to obtain point data, a cubic spline (De Boor, 2001) fitting method is employed to interpolate between ERA5 grid points, which, at the relatively fine grid scale of ERA5, is a technique with suitable performance levels (Balog et al., 2023; Tabor and Williams, 2010). As an example of the meteorological signals used in this thesis, Figure 3.3 shows meteorological variables extracted at Cambridge for the year 2000.



(a) Precipitation (b) Temperature

(c) Relative Humidity (d) U Component of Wind Speed

Fig. 3.2 Example climatic variables extracted from ERA5 and plotted over the full extent of the UK on the 1st of January, 2000, at midday, where precipitation is a surface variable and temperature, relative humidity, & wind speed are pressure variables extracted at a pressure level of 1000hPa.



(a) Precipitation (b) Temperature

(c) Relative Humidity (d) U Component of Wind Speed

Fig. 3.3 Example climatic variable signals extracted from ERA5 at the longitude and latitude for Cambridge for the year 2000, where precipitation is a surface variable and temperature, relative humidity, & wind speed are pressure variables extracted at a pressure level of 1000hPa.

Storms and Cyclones

The dataset provided by the European Severe Storms Laboratory as part of the European Severe Weather Database is extensive but all forms of extreme weather phenomena affecting the United Kingdom over the period of interest have an occurrence of approximately 5% (Dotzek et al., 2009). Having representative data is critical in ensuring high model performance and, whilst some of this activity affecting the UK is cyclonic, it might be insufficient. Therefore, although it might be more geographically consistent to use this database, it would be more expedient to expand the geographic scope to areas where specific types of extreme phenomena occur with far greater frequency.

By turning our attention to the North West Pacific Basin, we gain exposure to the tropical cyclones, or typhoons, and the high occurrence frequency for this one type of extreme phenomena. If warm water is the fuel that powers storms, as per our conceptual outline above, then the Pacific being a hotbed of storm activity necessarily follows due to the great expanse of water over which storms can form and the average surface ocean dynamics, exacerbated by the El Niño–Southern Oscillation driving interannual cycles of warmer than average surface water (Chan, 2000; Chen and Tam, 2010; Chia and Ropelewski, 2002; Wu et al., 2018). Using data from the Digital Typhoon Project, part of the Japanese National Institute of Informatics (Kitamoto, 2022), the proportion of days with active typhoons is 43% on average across all years of interest. Ergo, there is a richness of data in the North West Pacific, far exceeding that of the UK, with which to train the models and a near absence of dataset imbalance. As an aside, with regards to a weather system that might be considered extreme, we note that one might best consider events extreme not by their probability of occurrence but by their potential impact.

To clarify further, a tropical storm is a cyclone with wind speeds in excess of circa 60kmh^{-1} and a typhoon is one with wind speeds in excess of 120mh^{-1}, if occurring within the Western Pacific or North Indian oceanic basins; insofar as this work is concerned, given that typhoons are subsets of tropical storms that are, in turn, subsets of all cyclonic storms, we will refer to both tropical storms and typhoons as being storms.

In more detail, this dataset details the longitude and latitude for storms at 6 hourly timesteps along with the associated pressure; also provided is the maximum radius of storm wind, gale wind, and metrics such as the total power dissipation of the storm. In Figure 3.4, we have provided an example from the database towards the start of our period of interest, the Typhoon Tip, which ran from approximately 06:00 on the 4th of October 1979 to 06:00 on the 22nd of October 1979.

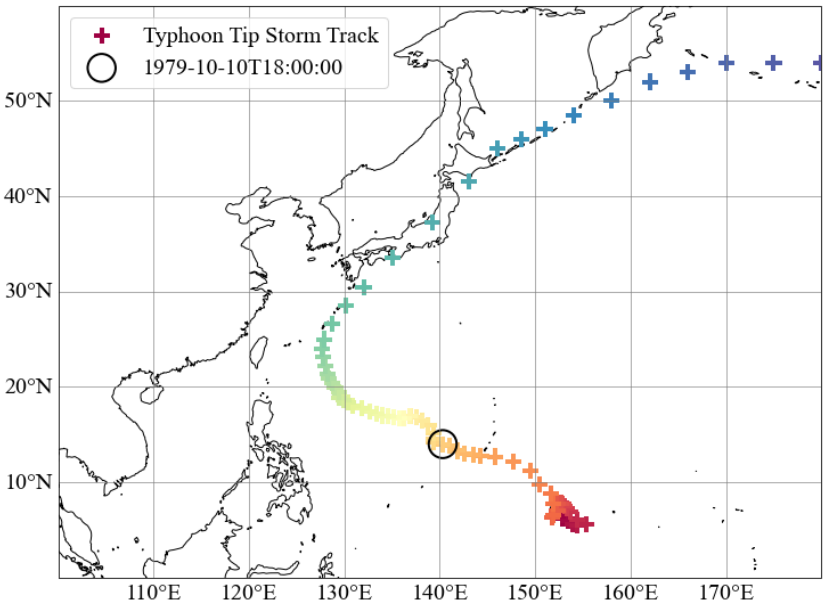
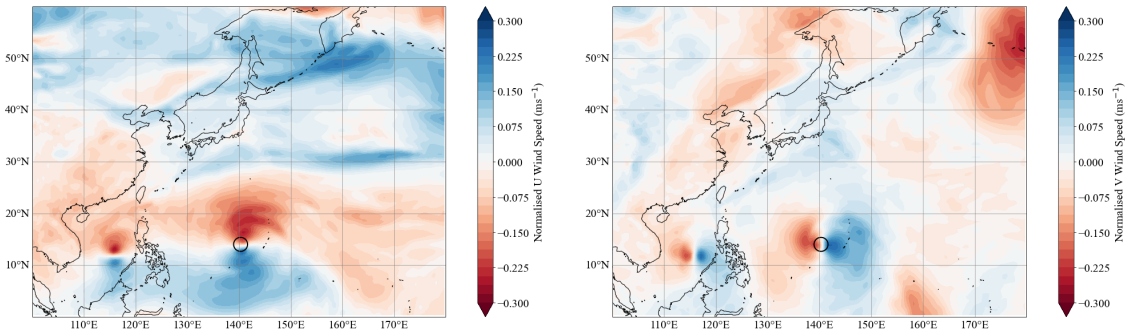


Fig. 3.4 Storm track for Typhoon TIP in the North West Pacific Basin, gradient coloured by time with 6 hour step and cyclogenesis at 06:00 on the 4th of October 1979, circa latitude 6.2° and longitude 152.9°, with reference point circled at 18:00 on the 10th of October 1979 for Figure 3.5.



(a) Contour plot of normalised U component of wind speed

(b) Contour plot of normalised V component of wind speed

Fig. 3.5 ERA5 wind speed data at 18:00 on the 10th of October 1979 in the two dimensions parallel to the surface of the earth at a pressure level of 750hPa, at which point the typhoons Sarah & Tip are active in the North West Pacific Basin, with the centre of typhoon Tip highlighted in both for comparison with Figure 3.4

The data provided through the database is supported by satellite imagery and radar; thus, the more interesting quality assurance lies in the comparison between the storm tracks and the ERA5 data. In Figure 3.5, we have compared an instance ERA5 data, plotted as the contour, with the reported location from the storm database to show the overlap between the two; the same instance is highlighted in the track in Figure 3.4. Across the majority of storms analysed in this way, the same agreement between what will comprise the data input, ERA5, and the data labels, data from the storm database, was found. Furthermore, through the components of wind speed, the structure of a typhoon is well represented, with both the eye and eye wall evident. Therefore, we have minimal concerns about these forming an input and output pairing.

3.2.2 Hydrological Data

Streamflow Data

The UK Centre for Ecology and Hydrology provides extensive records of hydrometric data from across the UK. There are 1,602 gauge station records available through the UK National River Flow Archive (NRFA) (UK Centre for Ecology & Hydrology, 2022), some of which are incomplete or at which measuring was concluded after a relatively short period, being years rather than decades. Ergo, a strategy was required for filtering out those which would not be appropriate for this study. Given that nonstationarity in the face of climate change is a key theme for the methodologies developed in this work, streamflow records that are of significant length and complete within that time period are of more use as part of the training set. Significant, in this case, would necessarily need to have overlap with the ERA5 datasets listed above. Thus, ideal candidates have complete temporal records between the years of 1979 to 2019 for either the preliminary modelling applications and for the generalising approach we wish to adopt. In terms of the single catchment modelling, the split between the training set and test set should not be done arbitrarily; by selecting the earlier part of a catchment's historical record for training and using the later part to test the model would enable analysis of whether or not the model is able to capture and represent non-stationarity, if present within the record of data.

If we consider an example streamflow dataset, such as for the River Aire at Oulton Lemonroyd, selected arbitrarily, then the gauged daily flow is shown in 3.6; this gauged daily flow is taken over the course of 24 hours and, therefore, includes and is heavily influenced by storm, or peak, flows, although the response is smoothed and not representative of the instantaneous peak flow. As has been noted in literature (Bartens and Haberlandt, 2021; Ding et al., 2015; Fill and Steiner, 2003), our choice here is governed by data availability

but a shortfall of using daily flow is that, depending on the intensity and duration of a meteorological event in combination with catchment characteristics, the instantaneous peak flow can exceed the maximum daily flow to an extent that is significant. In particular, in small catchments, where the time for runoff to reach the river, due to maximal distance from the furthest extents of the catchment to the river, is likewise small, the instantaneous peak flow diverges considerably from mean daily flow (Fill and Steiner, 2003; Fuller, 1914; Gray, 1973). Other factors, including the topography and the land use also result in changes to the instantaneous peak response (Canuti and Moisello, 1982). Thus, the impulse response in certain circumstances may be underrepresented. Whilst this is a significant issue if one were to use daily flow data alone in the adaptation to and mitigation of floods, there are methods to estimate peak flow from maximum daily flow (Bartens and Haberlandt, 2021; Ding et al., 2015). Moreover, our aim is to demonstrate the generalising capability of machine learning methods, so that were the data readily available for instantaneous peak flow, they might be more readily applied.

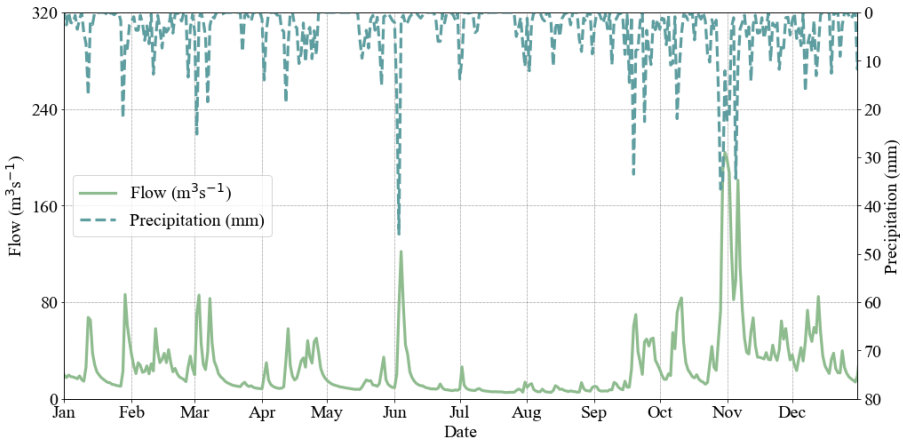


Fig. 3.6 Gauged streamflow for the River Aire measured at Oulton Lemonroyd over the year 2000 with precipitation shown on a secondary axis for the same time period, highlighting the congruence between peaks in precipitation and streamflow

Having said that, the extremes that we are especially concerned with here tend to have extended temporal properties that are, likewise, extreme. If we consider the 2007 Summer floods, which, at the time, experienced the highest volume of May-June precipitation on record (Marsh, 2008), the areas experiencing the worst flooding were affected by up to 24 hours of rainfall and the subsequent runoff and streamflow volumes that led to rivers bursting their banks were also of significant duration, thus better captured and represented in daily

flow data. One of those areas was Oxford (Marsh and Hannaford, 2007; The Environment Agency, 2007), for which we present the precipitation and streamflow record for the gauging station on the River Thames at Sutton Courtenay, downstream of Oxford and approximately 12km south, as shown in Figure 3.7.

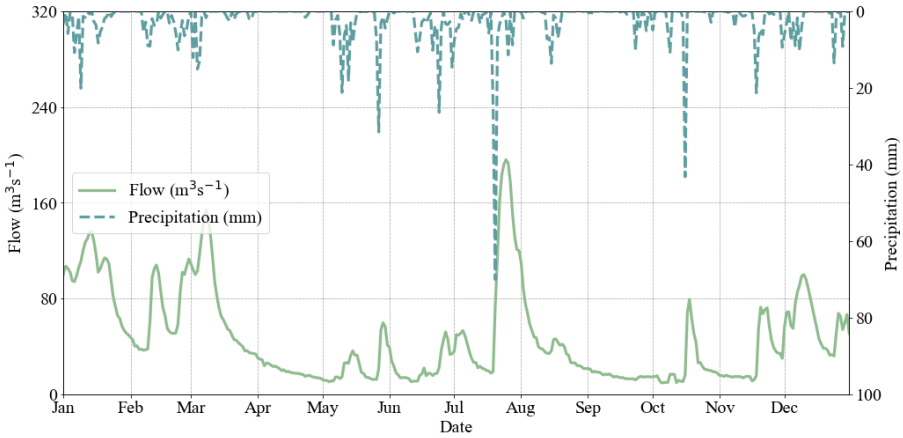


Fig. 3.7 Mean annual global surface temperature anomaly for each of the Shared Socioeconomic Pathway Scenarios

Catchment Descriptors

Catchment descriptors are variables that, as the name might imply, describe the physical properties of an individual catchment, in terms of the topography, geology, land use, et cetera; as described in the first section of this chapter, these variables have a direct impact on the internal flows of a catchment, thus affecting the streamflow and consequential flood risk.

The topography of a catchment is expressed through its elevation data, taken at the minimum, maximum, 10th, 50th, and 90th percentiles as measured relative to the UK Ordnance datum at Newlyn, Cornwall. This topographical information, including the catchment boundaries and areas comes from the Centre for Ecology and Hydrology's Integrated Hydrological Digital Terrain Model, derived using a combination of topographical, inland water, and coastline data (D. G. Morris and R. W. Flavin, 1990, 1994) and uses UK Ordnance Survey data (Ordnance Survey, 2017).

The second subset is that which describes land use, broken down into the proportion of the catchment with arable/horticultural, grassland, mountain/heath/bog, urban, or woodland covers. Land cover data has been assimilated through satellite imaging data covering the entirety of the United Kingdom in conjunction with a classification algorithm that used

spatial, spectral, and contextual characteristics to perform said classification of discretised elements of the domain (Fuller et al., 2002).

Finally, the subsurface geological data has been collected by the British Geological Survey (British Geological Survey, 2007). The British Geological Survey data has been compiled through extensive survey work done by a large team of geologists and supporting functions, using discrete direct measurement, such as borehole samples, along with interpolation methodologies that, more recently, includes three dimensional modelling to create full coverage maps (Entwisle, 2019; Kessler et al., 2009). The bedrock is broadly classified in terms of permeability with: high being that with highly productive fissured aquifers or aquifers with intergranular flow; moderate being that with locally important fissured aquifers or aquifers with intergranular flow; low being that with impermeable rock and negligible groundwater beyond soil depth; and mixed being that with concealed aquifers or those that are limited or only locally relevant.

The British Geological Survey and Ordnance Survey datasets are the most comprehensive of their kind covering the United Kingdom. Given that the UKCEH datasets are derivatives and also follow a rigorous process for data assimilation, then, as was the case with the streamflow data, we assume that minimal preprocessing of the data is required; thus, catchment descriptor data will be extracted from the respective data archives directly. As an example, if we consider the catchment for the River Aire at Oulton Lemonroyd the following table, Table 3.1, presents a comprehensive subset of the catchment descriptors that will be used within this thesis, primarily in Chapter 6. A further representation of one subset of the catchment data, being the topographical information, is presented in the left panel in Figure 3.8.

Table 3.1 Example set of catchment descriptors for the River Aire at Oulton Lemonroyd

Elevation		Land Cover		Geology	
Percentile	mAOD	Classification	%	Permeability	%
Minimum	13.10	Woodland	9.92	High	0.00
10th	75.20	Arable/Horticultural	4.24	Moderate	68.85
50th	175.40	Grassland	54.27	Very low	0.35
90th	344.90	Mountain/Heath/Bog	6.21	Mixed	30.81
Maximum	586.00	Urban extent	23.83		

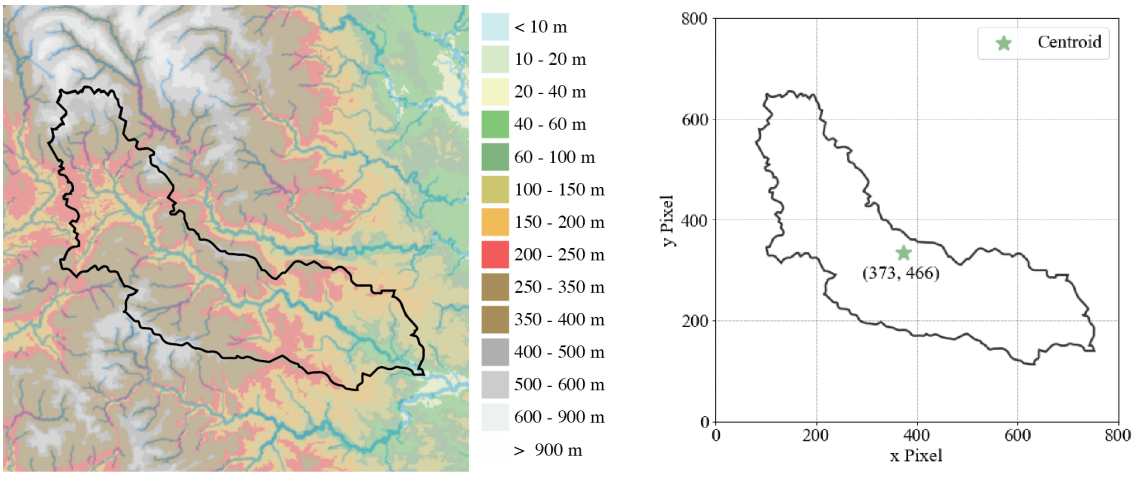
Catchment Centroids

For some of the modelling approaches adopted in this thesis, an important parameter of catchments is their centre; in this context, we define centre to mean the centroid of the catchment, being the geometric centre of a catchment's two dimensional surface, equivalent to its centre of mass for a congruent shape of uniform weight density. This parameter is required for interpolating or averaging meteorological variables between grid points.

Assuming that a catchment is a shape in two dimensions with area, A, its centroid, with coordinates $(\bar{x}, \bar{y})$, is given by taking first moments of area in both axes, with respect to some frame of reference, divided by the shape's area, as in Equation 3.2:

$$\bar{x} = \frac{\int_A x \, dA}{\int_A dA} \; ; \; \bar{y} = \frac{\int_A y \, dA}{\int_A dA} \tag{3.2}$$

For the purposes of this calculation, we will consider the catchment as a surface that exists in a flat, two-dimensional plane, effectively removing elevation and reducing the dimensionality of its boundary, B, such that $B \in \mathbb{R}^2$ rather than $B \in \mathbb{R}^3$ and treating that boundary as the continuous edge of an irregular shape with uniform density. This simplification is suitable as the primary forcing variable of interest is likely precipitation, as explored previously, and is treated as a surface variable in meteorological and climate modelling output.



(a) Catchment boundary depicted on a map of elevation from the NRFA

(b) Catchment centroid derived using a rasterised image

Fig. 3.8 Catchment boundary for the River Aire at Oulton Lemonroyd and associated centroid

Again, taking the arbitrary example of the Aire at Oulton Lemonroyd, the coordinates of the centroid of the catchment when projected into a pixel grid of dimensions 800 by 800 are $(373, 466)$. Overlaying this back in the latitudinal and longitudinal system returns a coordinate pair of 53.9° latitude and $-1.87°$ longitude.

3.2.3 General Circulation Model Data

For the final research chapter, Chapter 9, our objective is to force machine learning models using GCM data to make forward projections of the impact of climate change on streamflow, which can then be used in turn to make inferences about future flood risk. Over 100 models from 50 different modelling centres have been utilised for CMIP6, so a selection strategy is required. The objective should be to enable simpler downstream analysis whilst also minimising preprocessing; whilst the former is hopefully transparent, the latter is important because, ideally, we would prefer the simulation data to be as close to observations as possible and for as few transformations as possible to preserve the models' representation of physical processes and maintain spatiotemporal field consistency (Ehret et al., 2012). One option that might seemingly be easier to justify is to utilise all available models and combine them into an ensemble but the amount of preprocessing for this study would be somewhat prohibitive, so, at this juncture, we leave this as a future possibility and assume using a single model output is a reasonable strategy for predicting future changes under that model's rationale.

As to the selection criteria, having simulations for as many emissions scenarios as possible is key for comparative analysis and the model should have as high a resolution as possible to minimise the impact of spatial scaling discrepancy between prediction and observation (Challinor et al., 2009); ideally, the model output ought to be of the same temporal scale, utilising a Gregorian calendar, again to reduce the amount of data preprocessing for the reasons outlined above. These criteria serve to thin the herd out significantly and, from the remainder, the Max Planck Institute Earth System Model (MPI-ESM) complies with the requirements for this study; with an atmospheric resolution of 103km or 0.93°and an ocean resolution of 44km or 0.4°, the MPI-ESM offers high resolution and improved representation of atmospheric dynamics and ocean state, although improvement in long-standing biases is assumed to be modest (Gutjahr et al., 2019; Müller et al., 2018), where bias is the temporally invariant model error (Haerter et al., 2011).

As with the meteorological data described in the relevant section of this chapter, the spatial domain is the subset bounded by longitude of -8° and 4° and latitude of 48° and 60°. The temporal dataset in this instance we split into hindcasts and projections either side of the year 2014, as this is the final hindcast year from the GCMs. The variables remain the same at precipitation, temperature, wind speed, and relative humidity.

We also employ a second dataset with which to compare against the MPI-ESM model. The Met Office UK Climate Projections 2018 (UKCP18) are a set of projection tools that provide high-resolution spatially-coherent future climate projections at a range of scales (Met Office Hadley Centre, 2019); UKCP18 is driven by the 15 members of the Hadley Centre's ensemble model for 60km grid-scale global projections and 12 members models used for 12km regional projections, with the latter necessarily being regionally calibrated. This dataset, therefore, might be more appropriate for the domain and more specifically biased towards the UK.

Unfortunately, however, the UKCP18 dataset does have two significant drawbacks that prevent it from being the primary forcing dataset for the study. UKCP18 is driven by members of the previous incarnation of CMIP6, CMIP5, and therefore uses the Representative Concentration Pathway (RCP) scenario modelling that were used with the SSPs to create the scenarios for the latest IPCC Assessment Report (V. Masson-Delmotte et al., 2021). Furthermore, global projections are only provided for 2 RCP scenarios, those with the 2.6 Wm^{-2} and 8.5 Wm^{-2} radiative forcing, whilst the regional model has only been used to output projections based on RCP 8.5. Thus, in order to be able to continue to model the change between warming scenarios, the lower resolution dataset will be utilised for the aforementioned comparison.

Chapter 4

Establishing A Basis

With the motivation being the development of a system capable of generating predictions for areas with historically low data gathering activity, the temptation to adopt expansive and data rich modelling approaches should perhaps be ignored. Instead, the more prudent approach is to start with a most basic root. To construct this base machine learning framework for predicting streamflow, one must first begin with a hypothesis.

Thus, let us assume that streamflow on any given day for any unspecified river is a function of the rain that falls in the 24 hour period prior to the streamflow measurement being taken. This ought to appear reasonable as a starting point given that a flood hydrograph is used to assess the discharge response of a stream when subject to rainfall alone. Further emphasising simplicity, the models at this stage will be evaluated on a per river basin basis, rather than developing a framework that is able to generalise across geographies.

The study period will be set from 1979-2019 with a complete record over this period obtained for ERA at the time of writing and only those catchments in the NRFA database with a continuous record over that same period being selected for the catchment specific modelling, as per our data selection strategy detailed in Chapter 3. Out of all of the catchment records available, four have been randomly selected out of three different size brackets: two from the subset of catchments with catchment area, A, $> 2500km^2$, the River Severn observed at Haw Bridge and the Thames at Kingston, one from the subset with $1000km^2 < A \leq 2500km^2$, the Avon at Bathford, and one from the subset with $A \leq 1000km^2$, the Exe at Pixton.

A common, albeit imperfect, metric for assessing the performance or accuracy of models in hydrology is the use of the Nash Sutcliffe Efficiency (NSE) (Gupta et al., 2009; Nash and Sutcliffe, 1970). NSE is a normalisation of the Mean Squared Error (MSE), as shown in Equation 4.1, where y_i is the ith observed value, y_i' is the corresponding predicted value, and $\bar{y}$ is the average of all observed values.

$$NSE = 1 - \frac{\Sigma_{i=1}^{n}(y_i - y_i')^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.1}$$

The NSE will be used as the primary measure of the quality of fit for the basic models given its ubiquity and that it has been found to be a very potent hydrological objective function (D. N. Moriasi et al., 2007; Servat and Dezetter, 1991); whilst a full assessment of NSE is outside the scope of this body of work, additional metrics for the assessment of model performance will be used and brought in as might be appropriate. The NSE sits within the interval $[-\infty, 1]$, where an $NSE = 1$ represents the optimal value and anything ≤ 0 is no better than using the mean streamflow value; typically, $NSE \geq 0.5$ is considered acceptable and achieving an $NSE \approx 0.89$ would put a model on par with the median performance from a large collection of hydrological modelling studies (D. N. Moriasi et al., 2007).

4.1 Linear Models

4.1.1 Basic Linear Model

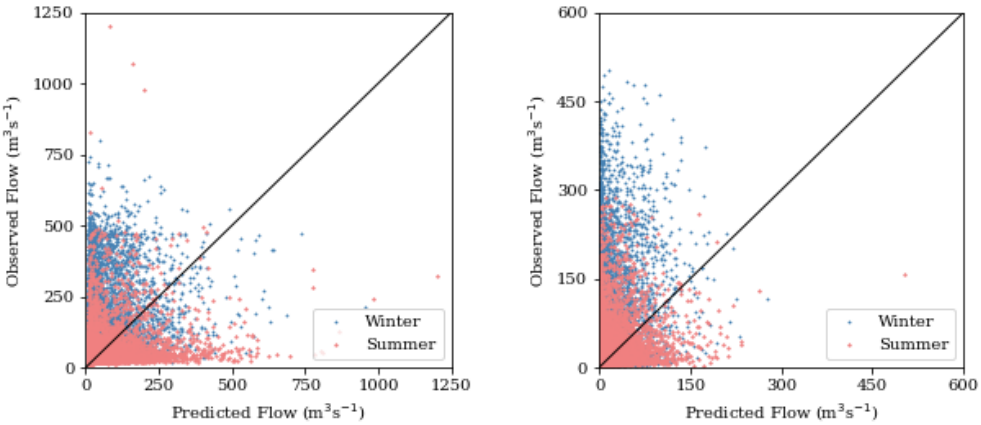
The assumption for the Basic Linear Model is that the daily streamflow in a river can be expressed as a linear function of the preceding twenty four hour's total average precipitation, denoted x, at the centroid of the catchment, as formalised in Chapter 3. The model is thus expressed in Equation 4.2.

$$f(x) = \phi \cdot x + \beta + \varepsilon \tag{4.2}$$

Where x is the cumulative rainfall over a single 24 hour period, ϕ is a user defined, rather than data-fitted, and universal coefficient that is then corrected for the range of the observed data on a per catchment basis, and β is the minimum baseline reading from the observed data, also on a per catchment basis, with some error term ε.

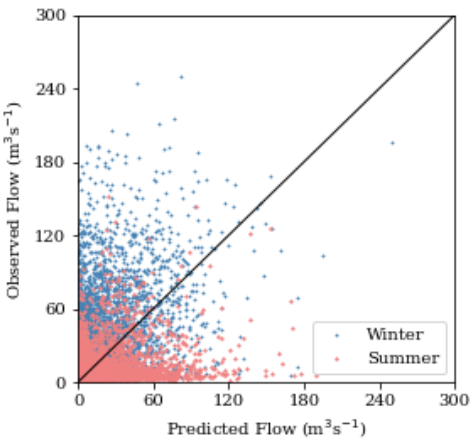
Table 4.1 Basic Linear Model prediction performance by catchment

Catchment	Metric						
	NSE	Flow Quantile (m^3s^{-1})					
		$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Severn at Haw Bridge	-0.530	35.6	12.4	66.4	23.4	135	70.1
Thames at Kingston	-0.428	12.7	1.73	35.1	4.70	79.9	19.7
Avon at Bathford	-0.226	4.51	0.364	9.26	2.83	25.0	15.8
Exe at Pixton	-0.473	1.38	0.382	2.46	1.22	5.70	4.96

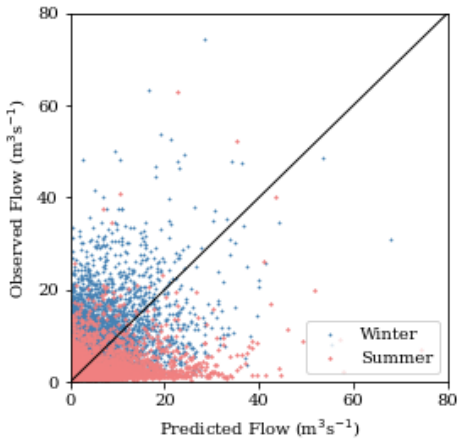


(a) Severn at Haw Bridge

(b) Thames at Kingston

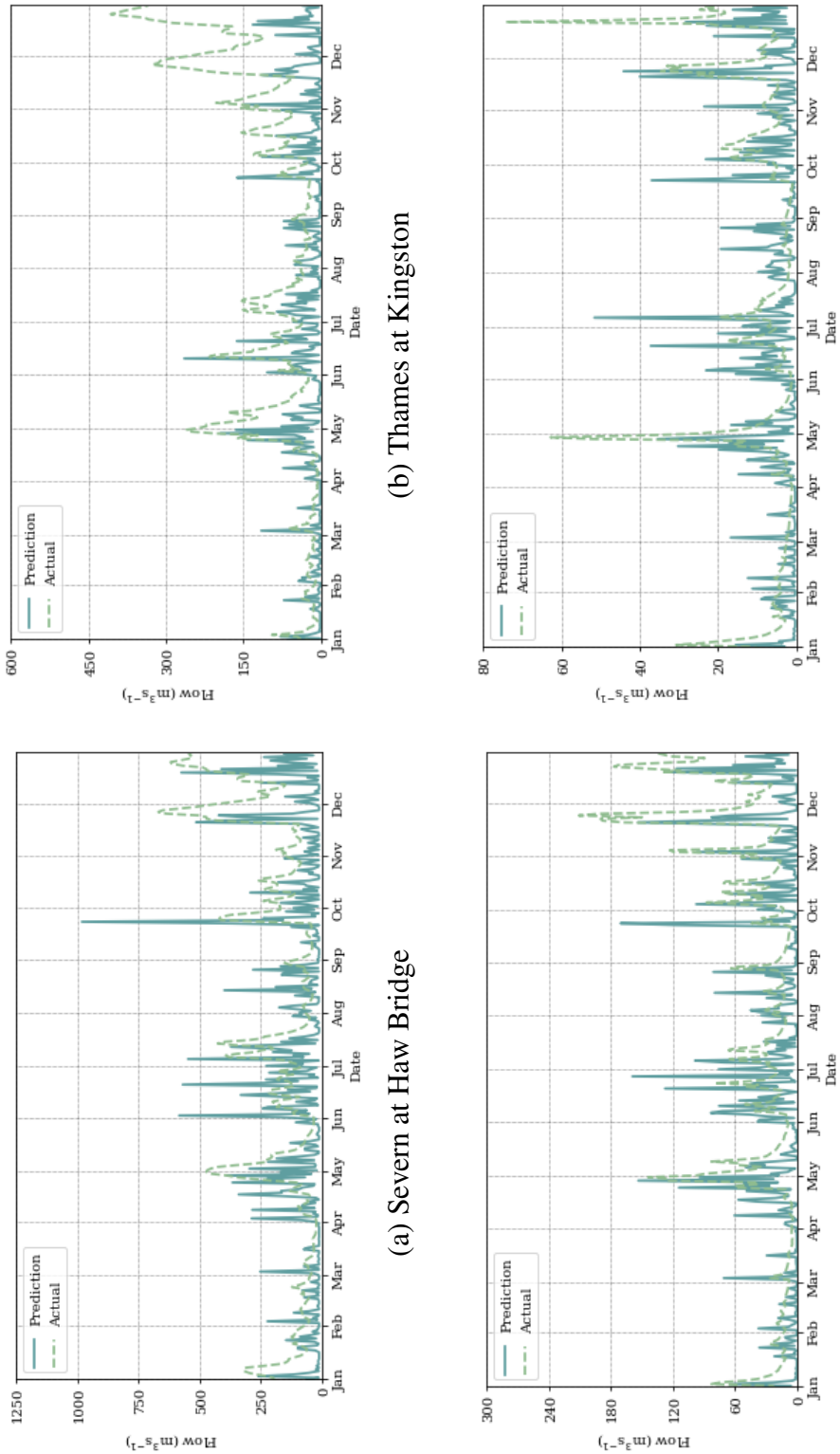


(c) Avon at Bathford



(d) Exe at Pixton

Fig. 4.1 Predictions against observations for the four test catchments using the Basic Linear Model



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 4.2 Comparative streamflow time series for the test set year, 2012, with predictions and observations using the Basic Linear Model

The result ought to not be surprising; the response according to this basic linear model is far noisier than the observed streamflow, as can be seen for all four catchments in Figure 4.2, and there is a lack of correlation between the predicted and observed values as directly plotted in Figure 4.1, with poor NSE values, all of which are shown in Table 4.1. However, one aspect of the results that ought to be considered is the similarity of the distribution of the results. In addition to the NSE, which is a dimensionless number, the table contains values for the 25th, 50th, and 75th quantiles for both the observations and predictions, denoted as $Q_{25:O}$, $Q_{50:O}$, and $Q_{75:O}$ and $Q_{25:P}$, $Q_{50:P}$, and $Q_{75:P}$, respectively. The results from this basic approach shouldn't be used as meaningful predictions for streamflow, given the poor fit as indicated by the NSE values, but there are facets that highlight the potential direction going forward; the distribution of the predicted values is not too dissimilar to that of the observed values and the catchment is responding accordingly to high volumes of rainfall. That the observed signal with respect to time is far smoother implies that there is a more significant temporal element to the catchment response than assumed with a univariate model.

4.1.2 Improved Linear Model

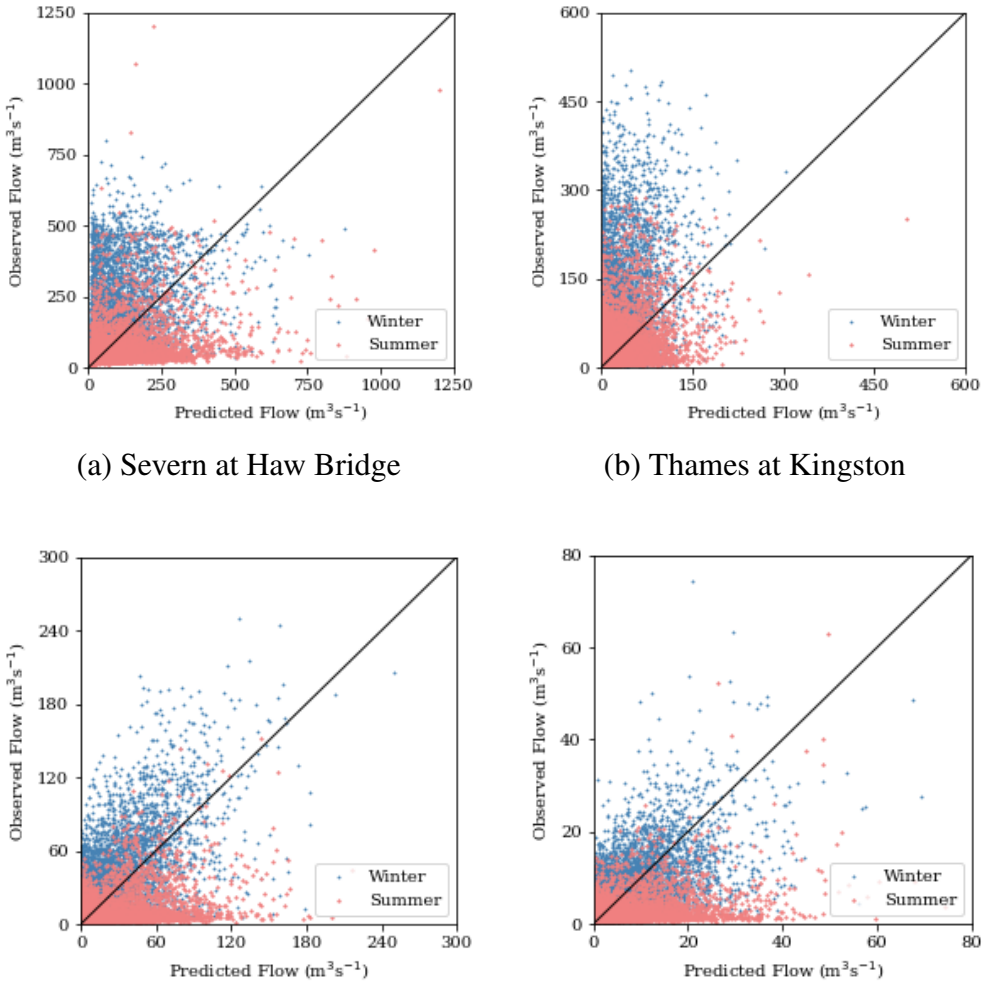
Improving upon the previous model, the most apparent change is to include an extra day of rainfall, with a view to both improving and smoothing the predicted response using the same four catchments and period of data. Thus, we present the similar Improved Linear Model, Equation 4.3, uses user defined coefficients, ϕ_1 and ϕ_2, to transform the two preceding and consecutive 24 hour periods of cumulative rainfall, x_1 and x_2, corrected for each catchment with minimum baseline β and error term ε.

$$f(x_1, x_2) = \phi_1 \cdot x_1 + \phi_2 \cdot x_2 + \beta + \varepsilon \tag{4.3}$$

The Improved Linear Model does, as the name would imply, offer some improvement over its predecessor, as shown in Table 4.2 and the plots of predictions versus observations in Figures 4.3 and 4.4; this improvement is limited, however, and the results imply that the Improved Linear Model is still not suitable for predicting streamflow, with the NSE for all catchments still being below 0. Having said that, the improvement is seen both in general fit and in the discrepancy between the distributions of the observed and predicted values and, although there is still being a significant amount of noise seen in the predicted signal with respect to time, further extending the rainfall record could add further improvement.

Table 4.2 Improved Linear Model prediction performance by catchment

Catchment	Metric						
	NSE	Flow Quantile (m^3s^{-1})					
		$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Severn at Haw Bridge	-0.335	35.6	17.1	66.4	44.7	135	114
Thames at Kingston	-0.217	12.7	3.00	35.1	10.9	79.9	35.7
Avon at Bathford	-0.094	4.51	1.34	9.26	8.00	25.0	28.6
Exe at Pixton	-0.893	1.38	0.696	2.46	2.96	5.70	8.92



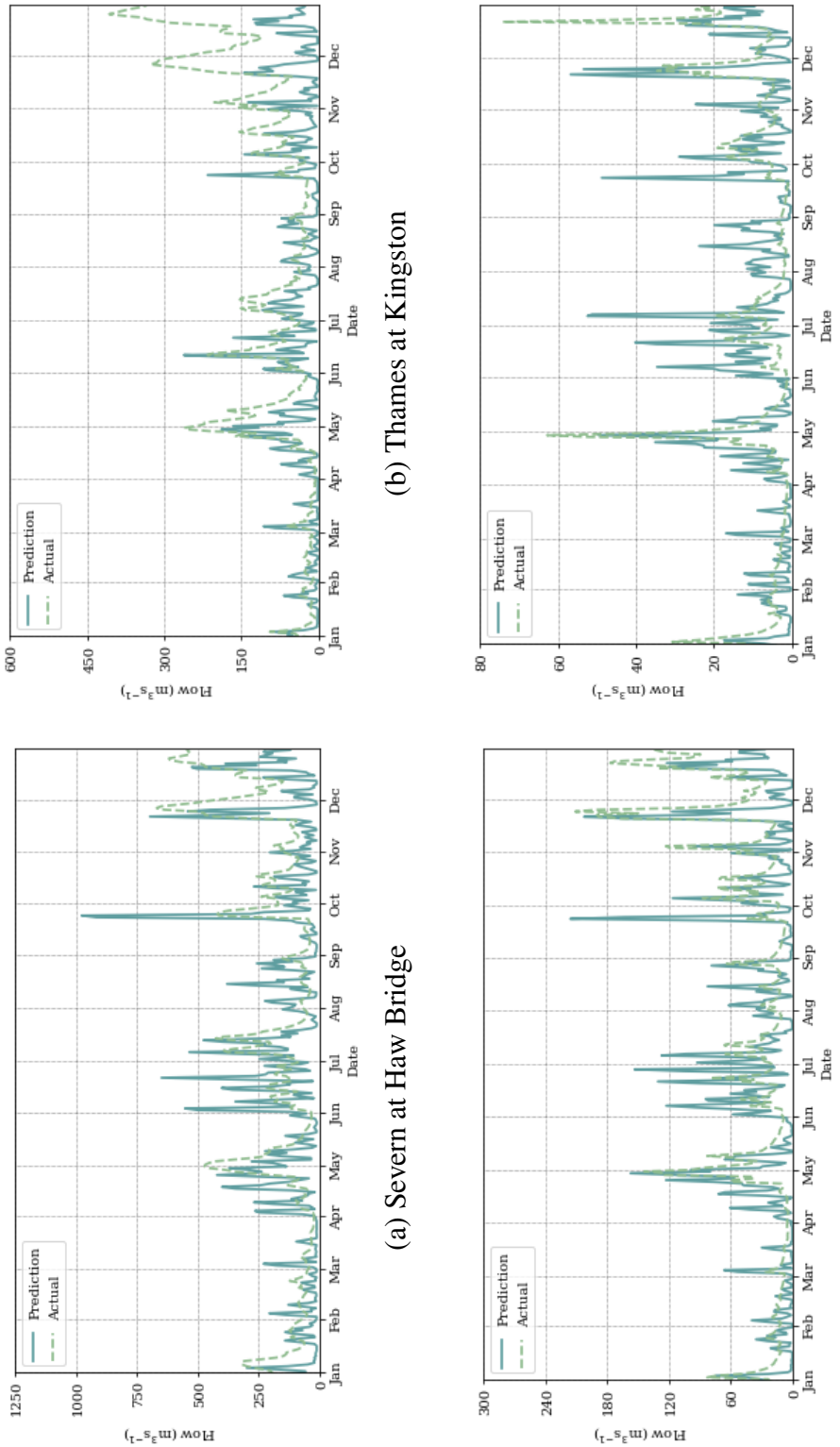
(a) Severn at Haw Bridge

(b) Thames at Kingston

(c) Avon at Bathford

(d) Exe at Pixton

Fig. 4.3 Predictions against observations for the four test catchments using the Improved Linear Model



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 4.4 Comparative streamflow time series for the test set year, 2012, with predictions and observations using the Improved Linear Model

4.1.3 General Linear Model

The General Linear Model marks a departure from the previous two approaches in that, although still linear, the parameters of the model are no longer user defined and are, instead, inferred from the data. As shown in Equation 4.4, streamflow is now expressed as a linear combination of $t = 7$ days of precipitation, such that $\mathbf{x} \in \mathbb{R}^t$ with parameters $\phi \in \mathbb{R}^t$ applied as weights to the t days of preceding, consecutive rainfall, with some error term ε.

$$f(\mathbf{x}) = \phi_1 \cdot x_1 + \phi_2 \cdot x_2 + ... + \phi_t \cdot x_t + \varepsilon \tag{4.4}$$

If this is then expressed in terms of the n number of data points, $\mathbf{Y}$ is the $(n \text{ x } 1)$ column vector of n responses, $\mathbf{X}$ is the $(n \text{ x } t)$ matrix of n inputs across t time steps, and Φ is the $(t \text{ x } 1)$ column vector of parameters to be inferred, s.t. $\Phi = [\phi_1, ..., \phi_t]^T$, as per Equation 4.5.

$$\mathbf{Y} = \mathbf{X} \cdot \Phi \tag{4.5}$$

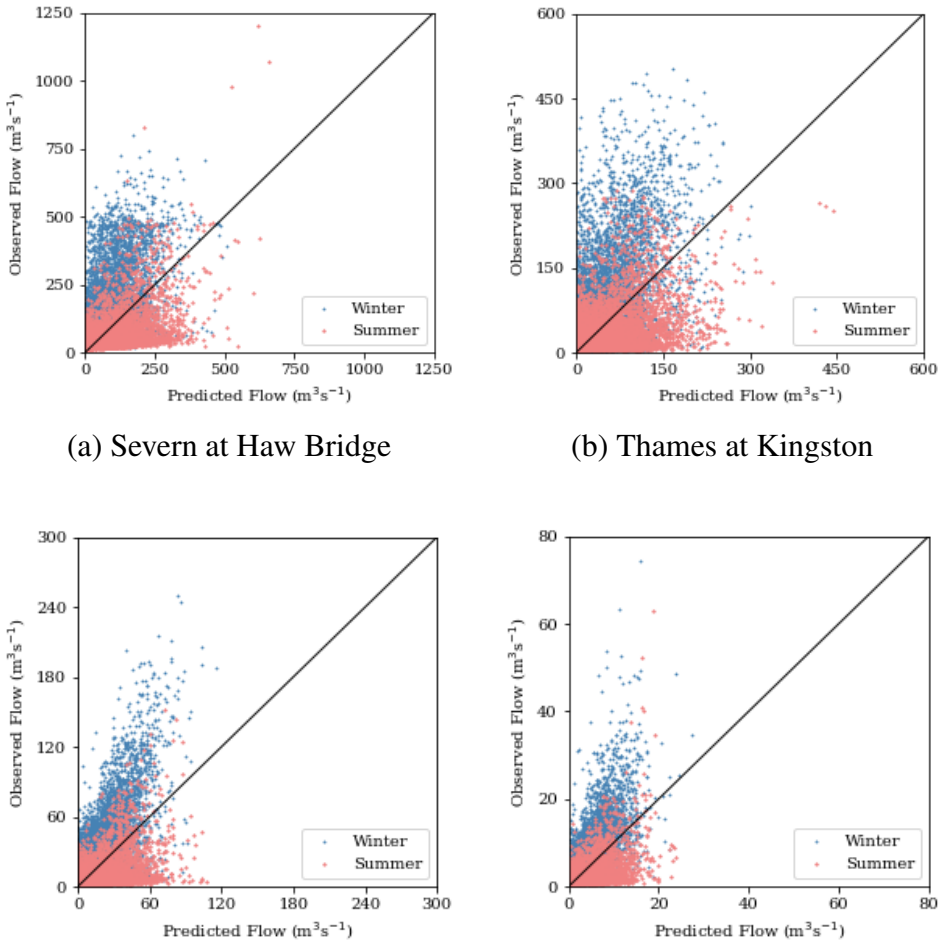
Fitting the model parameters is achieved through the manipulating the above into the normal equations and then solving for Φ, as per Equation 4.6.

$$\Phi = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{4.6}$$

Using the GLM, prediction distributions much more closely match those of the observations, with the NSE moving into positive territory. Whilst the overall improvement is sizeable, performance around extremes and seasonality is somewhat limited, as shown by the clear separation in predictive trend between summer and winter periods in Figure 4.5, with over- and under-prediction in summer and winter, respectively, and by the discrepancy between predictions and observations at the winter peaks in 4.6. Two fundamental aspects lie at the matter's crux: the input space is missing influential features or the model lacks the required expressive capacity. The latter issue will be tackled first with machine learning, moving from a linear to nonlinear model and gaining the ability to approximate a broader range of functions.

Table 4.3 General Linear Model prediction performance by catchment

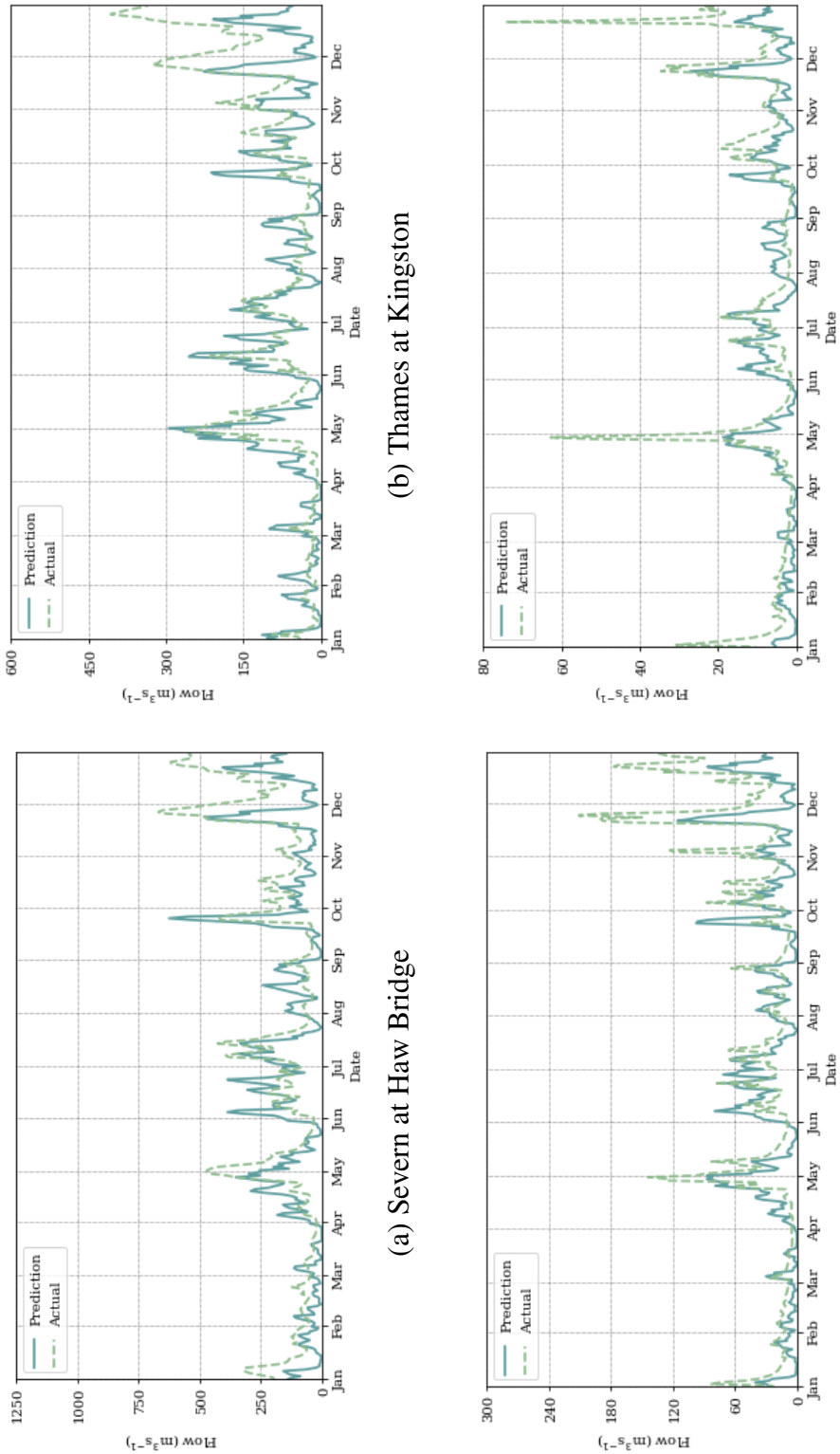
Catchment	Metric						
	NSE	Flow Quantile (m^3s^{-1})					
		$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Severn at Haw Bridge	0.053	35.6	23.7	66.4	61.3	135	117
Thames at Kingston	0.037	12.7	12.5	35.1	34.4	79.9	70.7
Avon at Bathford	0.241	4.51	3.67	9.26	11.4	25.0	25.2
Exe at Pixton	0.217	1.38	0.965	2.46	2.87	5.70	5.83



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 4.5 Predictions against observations for the four test catchments using the General Linear Model



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 4.6 Comparative streamflow time series for the test set year, 2012, with predictions and observations using the General Linear Model

4.2 Machine Learning

Machine learning is succinctly described by Mitchell as where "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience, E" (Mitchell, 2013); in this case, we already have both an idea of how to assess performance and the experience from which the program will learn, the remaining piece being to define the tasks.

Firstly, the application of machine learning in this work is restricted to the subdomain of supervised learning, with the goal being to learn a function that maps an input vector, $\mathbf{x}$, to the target vector, $\mathbf{y}$, given some dataset, $\mathfrak{D}$, of size n that can be expressed in terms of the input and output pairs, as in Equation 4.7:

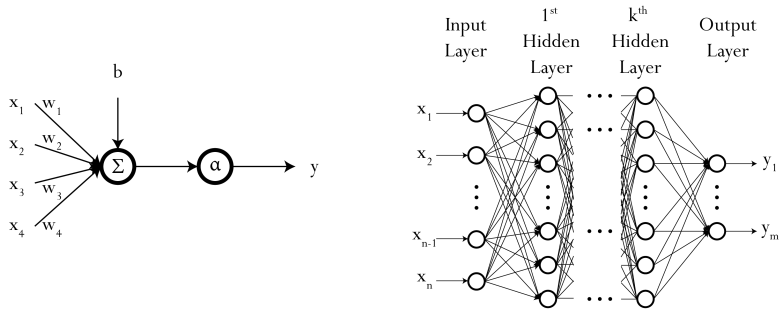
$$\mathfrak{D} = \left\{ (\mathbf{x}_i, \mathbf{y}_i) \right\}_{i=1}^{n} \tag{4.7}$$

Secondly, if machine learning is to be thought of as data driven decision making algorithms and data is half the equation (Lawrence, 2019), then the quality and composition of $\mathfrak{D}$ is paramount. Therefore, in keeping with a data first view of the problem, simpler, more explainable models will be used whilst efforts are focussed on the creation and curation of $\mathfrak{D}$. At this point, the first machine learning model of this thesis is introduced.

4.2.1 Artificial Neural Network Overview

The Artificial Neural Network (ANN) was inspired by the interconnected neurons in the brain and whilst, in spite of these biomimetic origins, any resemblance is likely on an abstract level the resulting model is suitable as a universal function approximator with widespread applications in nonlinear data modelling (MacKay, 2003; Wang, 2003). A Multi-Layer Perceptron (MLP) is composed of multiple, interconnected neuron units, each combining inputs linearly before an activation function is applied; depending on the choice for that activation function, nonlinearity can be introduced. Both a single neuron along with the structure for an arbitrary MLP are shown in Figure 4.7.

If we take a single neuron, then its output, y, in terms of its inputs, $\mathbf{x}$, and weights, ϕ, with added bias, β, and activation function, α, is expressed as in Equation 4.8:



(a) Single Neuron (b) MLP with l layers

Fig. 4.7 Graphical representation of a single neuron unit and an arbitrary neural network with l layers

$$y = \alpha\left(\Phi^T \cdot \mathbf{x} + \beta\right) \tag{4.8}$$

Using an MLP after the General Linear Model is a natural step given that if the activation function within an MLP is a linear function of the inputs, then the MLP, regardless of the number of hidden layers, simplifies to a linear model. Consider again Equation 4.8 and a linear activation function used for α, then the output, y_{ij}, of the ith node within a given layer, l_j, in terms of the inputs to the previous layer, l_{j-1}, is a linear combination of the inputs to the preceding layer again, l_{j-2}, as shown in Equation 4.9.

$$y_{ij} = \alpha\left(\Phi^T_{1,j-2} \cdot \mathbf{x}_{1,j-2} + \beta_{1,j-2} + \ldots + \Phi^T_{i,j-2} \cdot \mathbf{x}_{i,j-2} + \beta_{i,j-2}\right) \tag{4.9}$$

At this point of departure from linear models, suppose that the activation function, α, is some arbitrary non-linear function. Now suppose that all the weights, Φ, and biases, β, within the network are arbitrarily initialised at random values from an arbitrary interval, then upon making the forward pass through the network, a prediction y' is obtained. Given our arbitrary network parameters, then the prediction is likely to not be equal to the target, y; the expression of the error between the two is the cost function, $J(\Phi)$, expressed in terms of the set of all parameters, Φ, and can take the form of the Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), as per equation 4.10.

$$J(\Phi) = \left(\frac{1}{2n}\sum_{i=1}^{n}\left(y_i' - y_i\right)^2\right)^{\frac{1}{2}} \tag{4.10}$$

The objective, therefore, is to minimise the cost function, $J(\Phi)$, with respect to the set of its parameters, Φ, and all prediction and target pairs, y_i' and y_i. Minimisation of the cost function is done through gradient descent, an iterative calculus algorithm where each parameter, $\phi_{i,j}$, within the network is updated according to the sign and magnitude of the gradient to approach the values that correspond to a minima, as per equation 4.11.

$$\phi_{i,j} := \phi_{i,j} - \eta \cdot \frac{\partial J(\Phi)}{\partial \phi_{i,j}} \tag{4.11}$$

Where η corresponds to the learning rate hyperparameter that controls the rate of gradient descent. The Backpropagation algorithm (Rumelhart et al., 1986) enables the application of gradient descent, by propagating the derivatives backwards through the network after a forward pass through the chain rule of differentiation. Consider the parameters at the jth layer of a network with l layers, the update gradient for that layer is chained back through all previous layers and activation functions, $\alpha_l, \ldots, \alpha_j$, as in Equation 4.12.

$$\frac{\partial J(\Phi)}{\partial \Phi_j} := \frac{\partial J(\Phi)}{\partial \alpha_l} \cdot \frac{\partial(\alpha_l)}{\partial \Phi_{l-1}} \cdot \ldots \cdot \frac{\partial(\alpha_j)}{\partial \Phi_j} \tag{4.12}$$

Thus, the iterative learning loop is established, whereupon a forward pass through the network allows for the calculation of the loss and the use of backpropagation allows for parameter adjustment prior to the next forward pass, and so on until convergence.

4.2.2 MLP Application

The data needs to be split into subsets suitable for training, validating, and testing the model. Given that climate change is a non-stationary problem, splitting the data by year would enable quantification of the model's ability to fit changing trends, as have been observed in precipitation and flooding to date. Therefore, the period 1979 to 2009 will be used as the

training and validation set, further splitting through cross validation, whilst the period 2010 to 2019 will be used as the test set, with performance being assessed against the test set.

A relatively simple MLP structure will be adopted with relatively few layers, in this case just two hidden layers for a total of 4 layers including the input and output layers. The number of nodes within the first hidden layer will be set to half the number of inputs rounded up and the number of nodes within the second hidden layer will be set to half the number of nodes in the previous layer rounded up, e.g. for seven days of rainfall, the MLP structure would be $7 \rightarrow 4 \rightarrow 2 \rightarrow 1$.

In choosing an activation function, the conventional option that has seen extensive use and offers high performance is the Rectified Linear Unit, or ReLU, but a newer addition, the Swish or Sigmoid Linear Unit (SiLU) activation function, (Krizhevsky et al., 2012; Ramachandran et al., 2017). The SiLU function, defined in Equation 4.13, offers superior performance over the ReLU function, thought in part due it being a continous, non-monotonic function, whilst still being unbounded above and bounded below, like ReLU.

$$\alpha = \Phi(\mathbf{x}) \cdot \frac{1}{1 + e^{-\Phi(\mathbf{x})}} \tag{4.13}$$

Where the activation function, α, is expressed as in terms of the general linear transform of a node's inputs, $\Phi(\mathbf{x})$.

A common gradient descent optimisation algorithm, to be used for backpropagation, that has seen success over other procedures, including the base gradient descent algorithm, is the Adam algorithm, its name coming from adaptive moment estimation (Kingma and Ba, 2017; Ruder, 2017). Adam draws on both the adaptive gradient algorithm and root mean square propagation to include per parameter learning rates, η, adapted based on the biased corrected first, $\hat{m}_1$ in Equation 4.14, and second, $\hat{m}_2$ in Equation 4.15, moments of the gradients to be used in adjustment, Equation 4.16, during the backward step at iteration t. In practice, the hyperparameters, θ_1 and θ_2, work well with the default values as given in the originating paper and a small coefficient, ε, is introduced for numerical stability.

$$\hat{m}_1 = \frac{\theta_1 \cdot m_1 + (1 - \theta_1) * \nabla J(\phi)}{(1 - \theta_1^t)} \tag{4.14}$$

$$\hat{m}_2 = \frac{\theta_2 \cdot m_2 + (1-\theta_2) * \nabla J(\phi)^2}{(1-\theta_2^t)} \tag{4.15}$$

$$\phi := \phi - \eta \cdot \frac{\hat{m}_1}{\varepsilon + \sqrt{\hat{m}_2}} \tag{4.16}$$

Adam has been thought to give better training set error and begin to converge faster but, over a long enough training period, result in higher test set error (Wilson et al., 2017); whilst alternative optimisation strategies were investigated, such as using Adam for early training and then switching to stochastic gradient descent with gradient clipping, were looked at, they often yielded significant deterioration in test set performance and Adam was found to consistently give better results throughout this study. Given that this is not the main focus of the thesis, no further study is devoted to this aspect and Adam is used as the default optimisation strategy.

Finally, the inputs to the network will be rescaled such that they all fall within intervals that are of the same order of magnitude, using the normalisation procedure in Equation 4.17; each datapoint, x, is transformed by subtracting the mean, $\bar{x}$, and dividing by the range, between the maximum and minimum, x_{max} and x_{min} respectively, to give a normalised datapoint, x_{norm}:

$$x_{norm} = \frac{x - \bar{x}}{x_{max} - x_{min}} \tag{4.17}$$

Table 4.4 Shallow MLP prediction performance by catchment

Catchment	Metric						
	NSE	Flow Quantile $(m^3 s^{-1})$					
		$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Severn at Haw Bridge	0.297	36.8	61.7	64.8	92.4	134	143
Thames at Kingston	0.237	11.7	36.6	29.9	50.8	79.5	75.6
Avon at Bathford	0.348	4.24	10.2	8.53	15.2	22.8	23.4
Exe at Pixton	0.345	1.26	2.40	2.30	3.83	5.44	5.94

Comparing the results from this shallow MLP, shown in Table 4.4, there is a significant improvement in the model fit, with the NSE average across the four catchments increasing from 0.137 to 0.306 and beginning to approach the minimum acceptable benchmark laid out at the beginning of the chapter; the model evidently has greater expressive capacity and allows for more complex function fitting when using the same input space. Lack of model accuracy with regards to seasonality continues to be a problem, though, with weaker performance around winter; so additional variables that reflect the change in a hydrological system or its inputs with respect to season change are likely needed. Reflecting on the shortcomings of both the General Linear Model and the MLP, the combination of more suitable feature engineering alongside the benefits that the MLP offers should likely result in a model that is suitable for streamflow prediction.

4.3 Feature Engineering

The issue with the models' performance so far is that the full set of inputs and outputs, the hydrological state, and some of the internal processes of a catchment are not being accurately accounted for; at this juncture, the original hypothesis, that streamflow can be predicted accurately, in terms of the resulting NSE, solely using rainfall, should be considered invalid. Instead, if we refer back to Chapter 3 and consider the physical processes present within a catchment, a more appropriate hypothesis in terms of the climatic variables that can drive a catchment's streamflow response can be developed.

Rather than discounting the MLP, this is the more apt response to its underperforming, given that the work reviewed in Chapter 2 highlighted models were capable of achieving an $NSE > 0.9$, though this is often achieved through providing a number of previous daily streamflow measurements or predictions in order to predict streamflow at the next time step (Aichouri et al., 2015; Ali and Shahbaz, 2020). The use of MLPs clearly has potential to deliver outstanding performance, although the use of said previous predictions does, in a sense, go against an objective here and is not suitable for generating predictions for ungauged catchments (Besaw et al., 2010). Removal of previous days worth of streamflow within these studies resulted in significant lowering of performance but we might, instead, be able to internalise a river's previous conditions and achieve the same start of the art performance. Ergo, a superior, gauging-ignorant methodology is required to account for antecedent conditions in the catchment.

To restate, the total mass of water entering, leaving, and stored with the catchment must be equal to zero, as per Equation 4.18, expressed, for a time period of length T, in terms of the cumulative precipitation landing on the catchment, $\int_{t=0}^{T} \Psi_t dt$, the water leaving the

catchment through evapotranspiration, $\int_{t=0}^{T} E_t dt$, the change in water storage, ΔH, and the streamflow egress via the target river, $\int_{t=0}^{T} Y_t$.

$$0 = \int_{t=0}^{T} \Psi_t dt - \int_{t=0}^{T} Y_t dt - \int_{t=0}^{T} E_t dt - \Delta H \tag{4.18}$$

$$\therefore \int_{t=0}^{T} Y_t dt = \int_{t=0}^{T} \Psi_t dt - \int_{t=0}^{T} E_t dt - \Delta H$$

Although the change in storage, ΔH, is an essential part of this continuity equation, it can potentially be viewed, and indeed will be, as an internal part of the hydrological circuitry of a catchment. Therefore, if the model also implicitly represents the internal processes affecting water storage and attenuation, then the features that affect streamflow, our target variable, are assumed to be climatic only.

Our experimental setup remains the same with regards to subsetting the data, the catchments used, the performance metrics, and the broad structure of the MLP, whilst the input space will be changed to reflect inclusion of additional climatic variables and also further increasing the record of said variables. The climatic variables to be added are temperature, resultant wind speed, and humidity, as per our discussion in Chapter 3, and all of which will be added simultaneously, whilst further extending the record of these variables from 7 days to 28 days, as shown in Table 4.5.

Table 4.5 MLP model performance using increasing length of climatic variable record

Catchment	NSE for t number of days			
	$t = 7$	$t = 14$	$t = 21$	$t = 28$
Severn at Haw Bridge	0.621	0.675	0.712	0.743
Thames at Kingston	0.556	0.582	0.593	0.629
Avon at Bathford	0.678	0.696	0.727	0.752
Exe at Pixton	0.641	0.688	0.691	0.701
Average	0.624	0.660	0.681	0.706

Upon utilising the expanded climatic variable feature set, the improvement in model performance at the 7 day mark, is significant and now presents industrially significant performance. Further extending the climatic record from 7 to 28 days offers steady increases in model performance, as more information about the antecedent conditions within a catchment made available to the model is increased. Examining the relative impact of each of these additional climatic variables necessitates a comparative assessment of their effect on the predictive performance. Rather than presenting all possible permutations of the climatic input variables with respect to the number of days, increasing from 7 to 28, for all four catchments, which corresponds to approximately 1,200 different model evaluations, an elimination study will be presented instead. For the long climatic record, $t = 28$ days, the full set of input variables will be used and then the each one of the types of climate variable will be withheld, for example temperature would be omitted with respective the drop in performance recorded and compared against the same for each of the other variables. The effect of removing each subset of climatic variables is shown in Table 4.6.

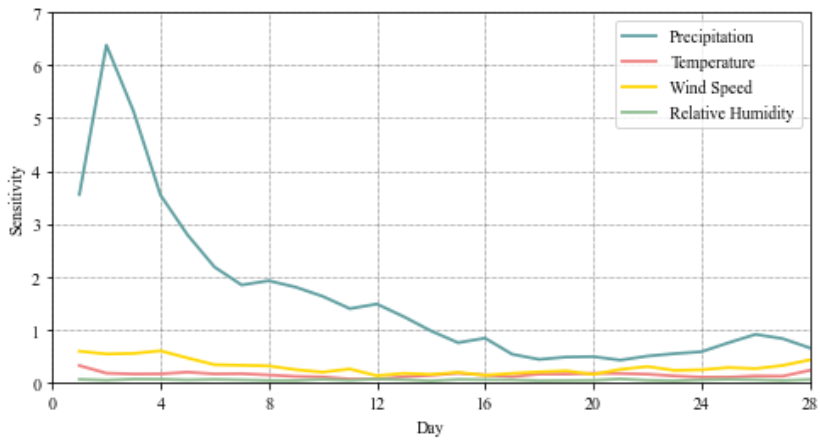
Table 4.6 Feature relevance to predictive capability assessment via elimination

Catchment	Dropped input variable NSE				
	None	Precipitation	Temperature	Windspeed	Humidity
Severn at Haw Bridge	0.743	-0.073	0.523	0.425	0.658
Thames at Kingston	0.629	-0.125	0.389	0.426	0.574
Avon at Bathford	0.752	0.187	0.689	0.719	0.728
Exe at Pixton	0.701	0.420	0.615	0.582	0.660
Average	0.706	0.102	0.554	0.538	0.655

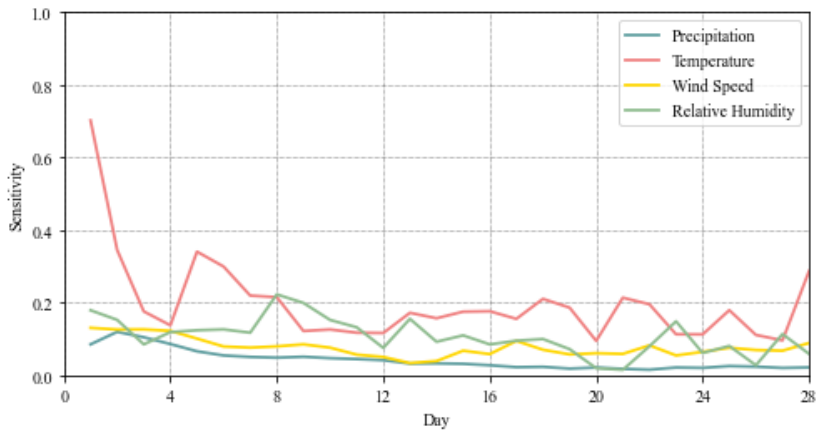
A further sensitivity analysis is provided through a perturbation study, (Scardi M, 1996), where all variables are set to their respective means, $(\bar{x}_1, ..., \bar{x}_n)$ to provide a baseline response, $\bar{y}_\mu$; each one of the variables, x_i, can then be set to either their maximum, as shown in Equation 4.19, or minimum individually with the network response normalised via the baseline response to provide a sensitivity, S, to that variable, respectively.

$$S = \frac{f(\bar{x}_1, ..., \bar{x}_{i-1}, x_i^{(max)}, \bar{x}_{i+1}, ..., \bar{x}_n)}{\bar{y}_\mu} \tag{4.19}$$

The network sensitivity to the maximum and minimum of each variable, averaged over the four catchments, is shown in Figure 4.8.



(a) Network response to variable maximum



(b) Network response to variable minimum

Fig. 4.8 Average network sensitivity to each variable at time, t, for high input signal, top, and low input signal, bottom

Streamflow is a heavily skewed distribution, best described by a Gamma distribution, so the variables that more drive increased volumes of water within the catchment and subsequent outflows ought to be those which the network is most sensitive to. From both the elimination and perturbation studies, the most important variables for prediction accuracy

are precipitation variables, with the network being particularly sensitive to high-volume precipitation within the period up to two weeks out from the predicted the stream flow, as shown in Figure 4.8. The network is then most sensitive to temperature, specifically low temperatures.

4.3.1 Soil Moisture & Proxy

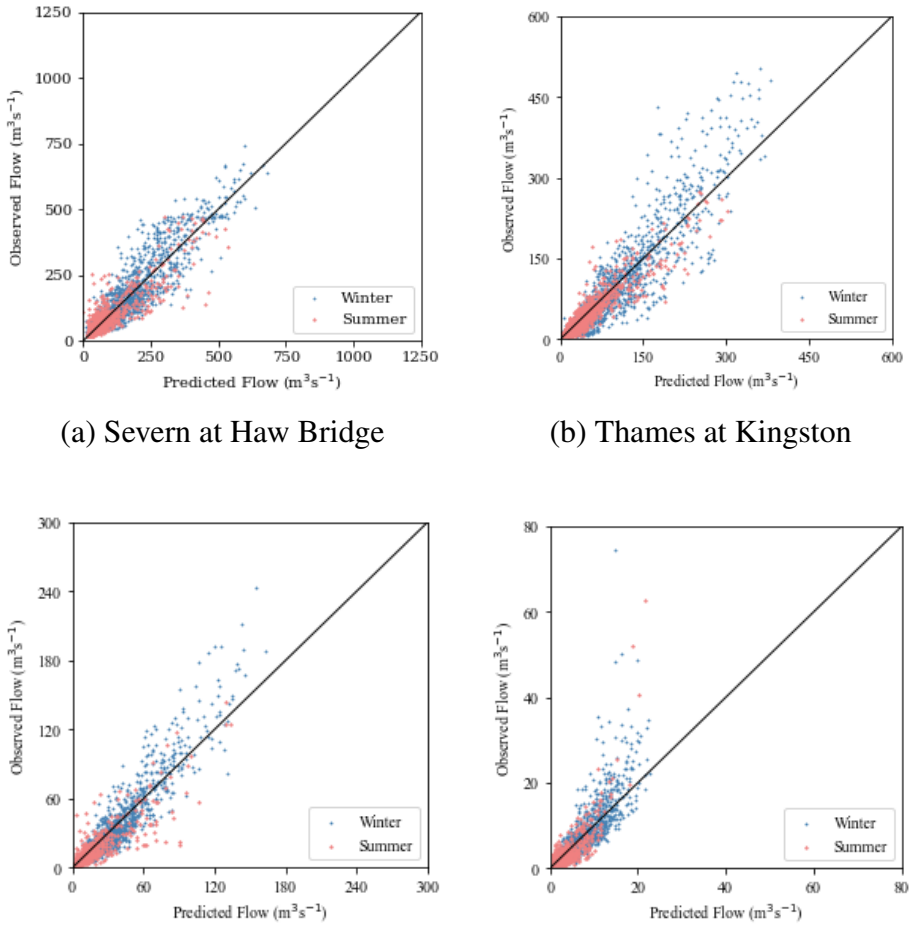
In essence, the input space investigated in the previous section on feature engineering is likely expansive enough to enable the simple ANN to capture the antecedent conditions and account for water attenuation, storage, and other internal processes of a catchment. However, the number of input variables in that final model is high, with four weeks worth of four primary variables such that $\mathbf{x} \in \mathbb{R}^{112}$. In the interests of being efficient with the model structure, being able to reduce the dimensionality of the input space is desirable and ensure that the model complexity is reduced as much as possible whilst retaining a high level of accuracy.

Thus, soil moisture variables will be introduced both alongside and in place of the extensive climatic record already supplied. Using the same experiment design, in terms of network architecture, test and training sets, etc., the comparative performance on the test set for our sample catchments can be determined for the four week climatic record input space with and without the addition of the soil moisture variables and for the original input space, of just seven days worth of rainfall. The data for soil moisture is taken at four depth levels up to a maximum of 2m depth below surface.

Table 4.7 MLP model performance using climatic variable and soil moisture input

Catchment	NSE for soil moisture based feature sets				
	$t = 7_{Precipitation}$	$t = 7$	$t = 14$	$t = 21$	$t = 28$
Severn at Haw Bridge	0.811	0.842	0.850	0.837	0.857
Thames at Kingston	0.828	0.846	0.852	0.858	0.851
Avon at Bathford	0.831	0.866	0.869	0.868	0.866
Exe at Pixton	0.677	0.710	0.716	0.714	0.706
Average	0.787	0.816	0.822	0.819	0.820

Inclusion of soil moisture in the input space enables shortening of the climatic variable record and offers a further increase in performance, as shown in Table 4.7 and Figures 4.9 and 4.10, with the mean NSE across the four catchments rising from 0.706 for all variables without soil moisture at $t = 28$ to 0.816 for all variables with soil moisture at $t = 7$. Even



(a) Severn at Haw Bridge

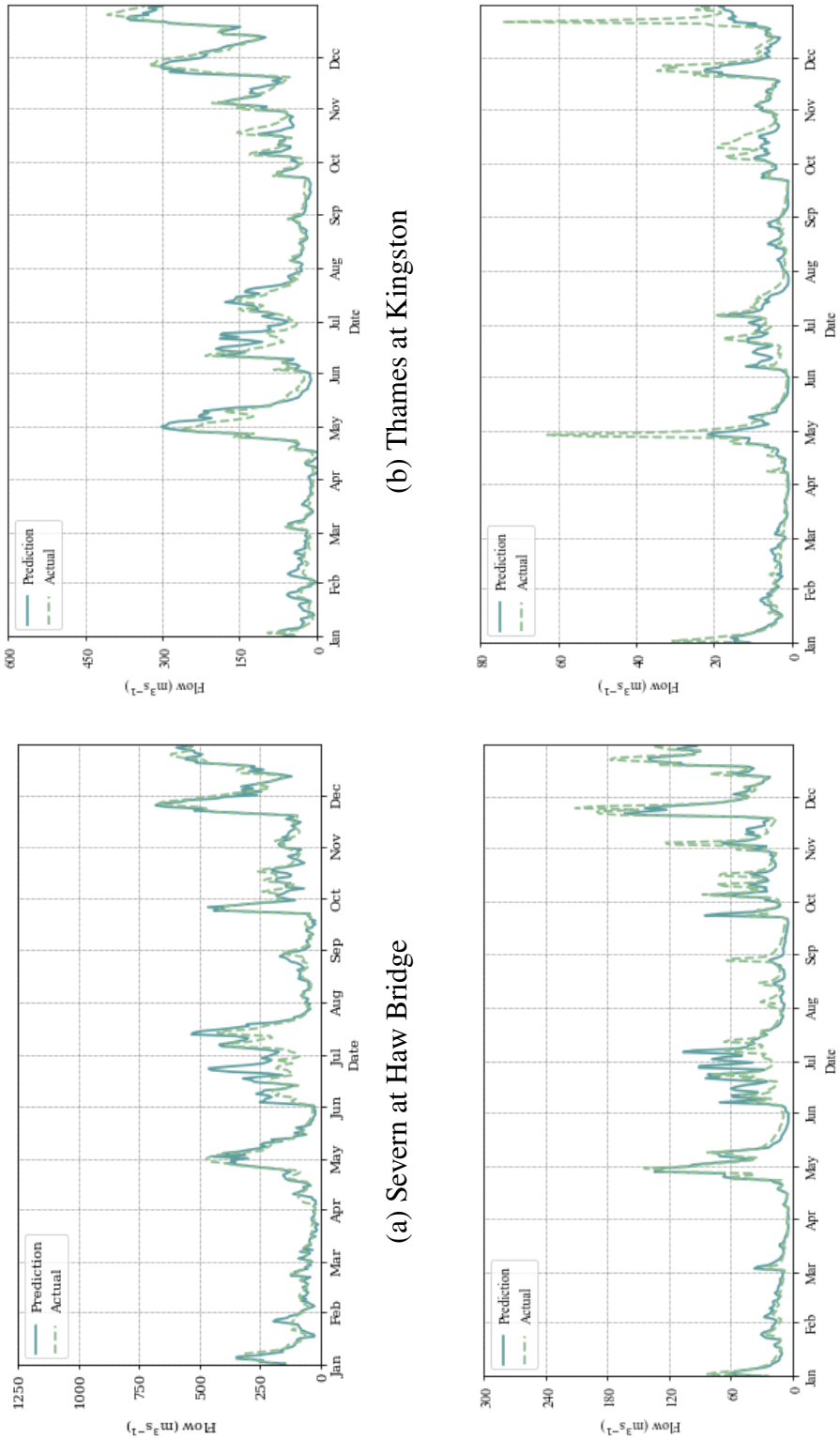
(b) Thames at Kingston

(c) Avon at Bathford

(d) Exe at Pixton

Fig. 4.9 Predictions against observations for the four test catchments using the soil moisture based MLP with $t = 14$

when used in conjunction with rainfall alone, it serves to better model performance over use of the other climatic variables used. Beyond being a direct measure of the antecedent conditions within a catchment and therefore being a substitute for longer climatic variable record length, soil moisture is also likely better representing internal processes, such as percolation and base flow.



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 4.10 Comparative streamflow time series for the test set year, 2012, with predictions and observations using the soil moisture based MLP with $t = 14$

Whilst the input space that includes soil moisture is smaller in terms of dimensionality and more efficiently captures the hydrological state of a catchment, perhaps a more interesting approach, one in keeping with using climatic variables as the only input variables, would be to develop a proxy climatic variable set that still represents the hydrological state of a catchment and is of equal or similar dimensionality.

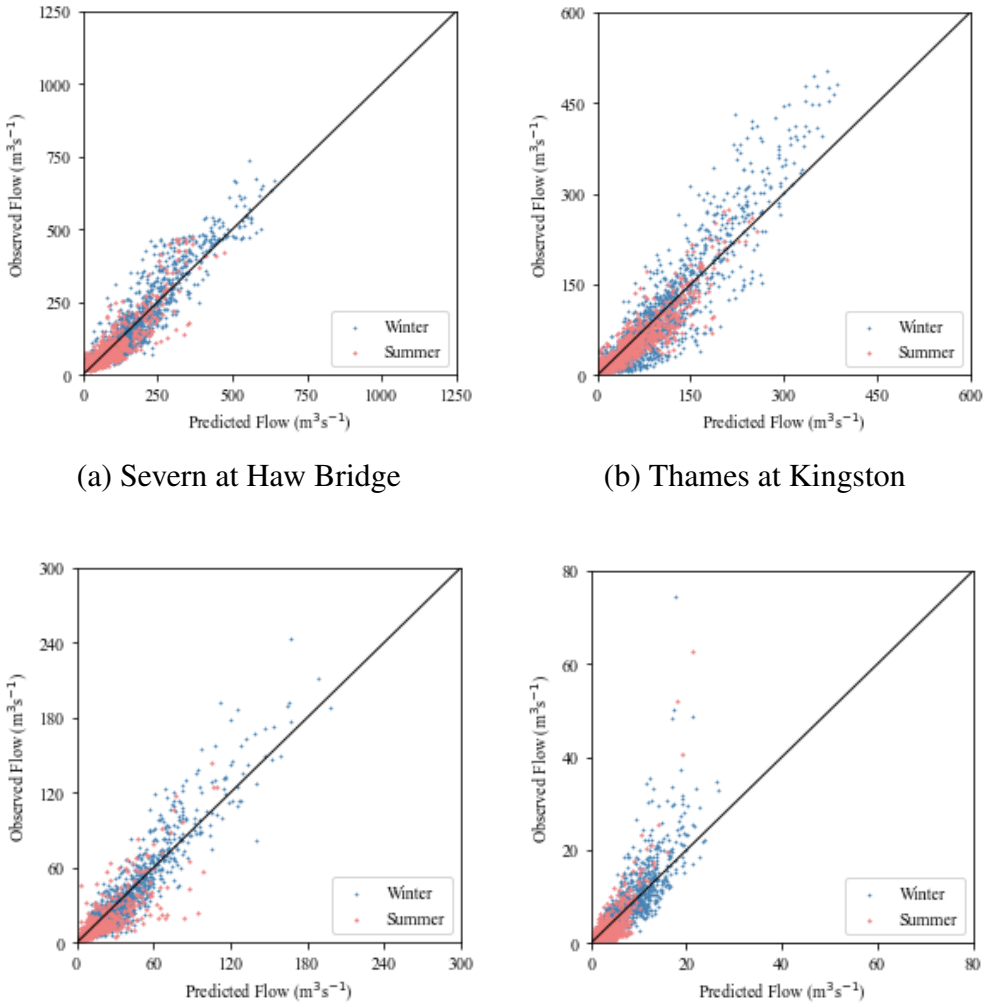
If the subsurface flows and physical structure of a catchment are treated as a subsystem that provides capacitance and storage, then a suitable analogue could be that of a function to smooth climatic input or a statistical representation of prior climate. Moving averages for the most impactful variables precipitation and temperature, as discovered when examining the sensitivity and response of the network, are taken for 1, 3, and 6 months and input into the network in place of soil moisture as antecedent proxies, with the resulting performance shown in Table 4.8 and Figures 4.11 and 4.12.

Table 4.8 MLP model performance using climatic variable and antecedent proxy input

Catchment	NSE for antecedent proxy based feature sets				
	$t = 7_{Precipitation}$	$t = 7$	$t = 14$	$t = 21$	$t = 28$
Severn at Haw Bridge	0.805	0.864	0.879	0.874	0.873
Thames at Kingston	0.843	0.854	0.857	0.853	0.851
Avon at Bathford	0.868	0.882	0.875	0.875	0.870
Exe at Pixton	0.668	0.700	0.710	0.707	0.709
Average	0.796	0.825	0.830	0.827	0.826

By using our proxy variables for antecedent conditions, the average performance across all catchments is further increased at all lengths of climatic variable record, although performance for the Exe at Pixton is marginally lower; notably, the discrepancy between some peaks in 2012, evidenced in Figure 4.12, remain and this issue has persisted across all approaches explored here. Further investigation of this anomaly is provided in Chapter 3.

These statistical measures serve both as a proxy to antecedent conditions but also act as a form of dimensionality reduction, compressing records of up to 180 days in length, whilst being highly interpretable and computationally efficient and easy to implement. Ergo, one can conclude, at this point, to have arrived at a feature space that achieves the required objectives of having no internal measurements, enabling the input space to contain only climatic variables and obtain a similar dimensionality, is explainable, and obtains the best performance thus far.



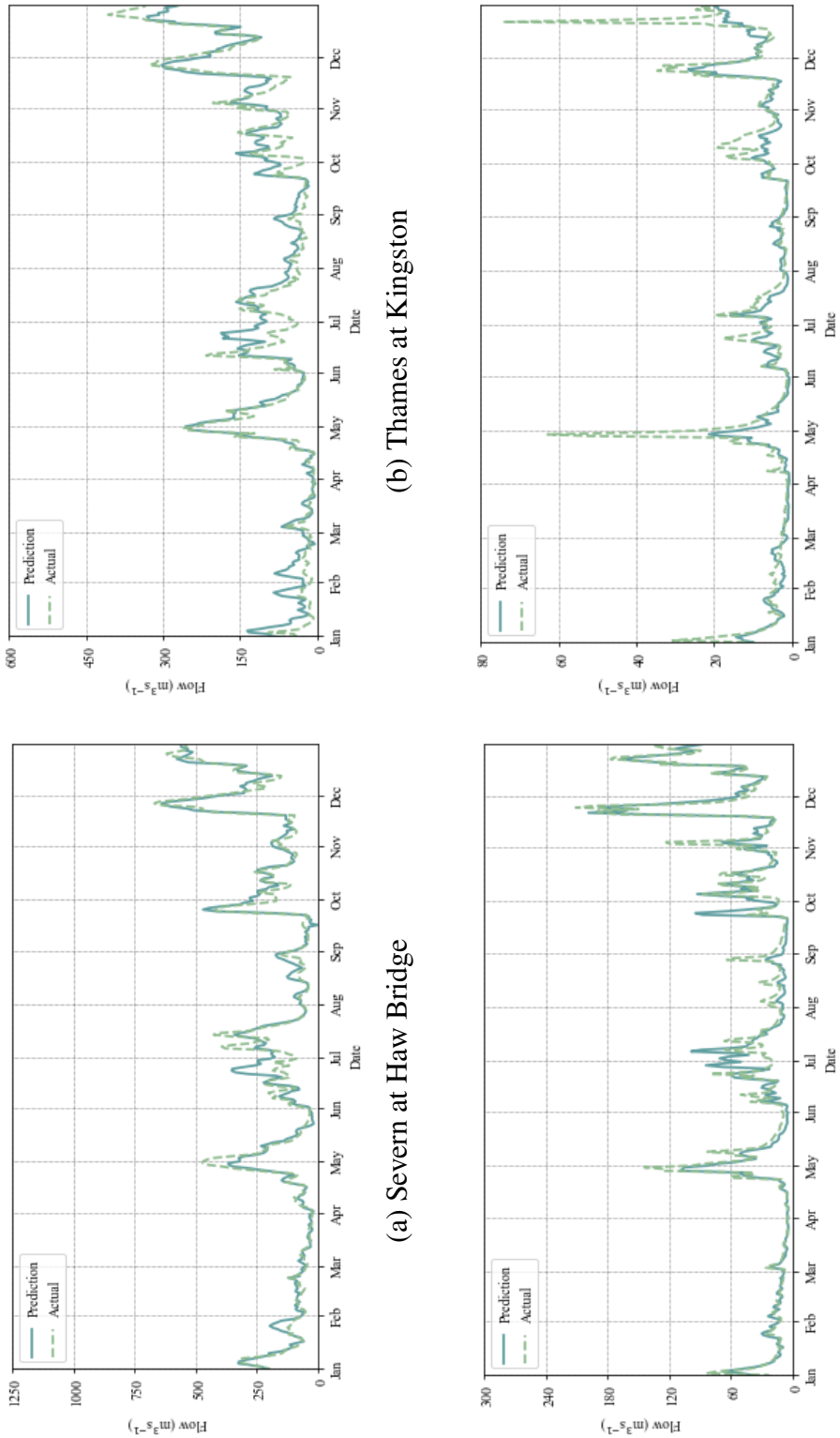
(a) Severn at Haw Bridge

(b) Thames at Kingston

(c) Avon at Bathford

(d) Exe at Pixton

Fig. 4.11 Predictions against observations for the four test catchments using the antecedent proxy based MLP with $t = 14$



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 4.12 Comparative streamflow time series for the test set year, 2012, with predictions and observations using the antecedent proxy based MLP with $t = 14$

4.4 A Model Foundation

The approach in this chapter has been to build a suitable model from the ground up; whilst the simple linear models used as a starting point do not offer much, either in terms of performance or novelty, the route taken has enabled the development of a lean, data-efficient model that does offer acceptable performance. The key output here is a machine learning approach with the justification that, as a nonlinear empirical model, it is possible to model streamflow to a relatively high degree of accuracy using only meteorological variables.

Further in part due to the approach taken here, it has been possible to show that simple, statistical measures of past climate can be used to approximate antecedent conditions within a catchment. This, therefore, eliminates the need for using lengthy climatic records in the input space and also serves as a suitable replacement for soil moisture variables and possibly other internalised variables. Concretely, we have demonstrated that internal variables can be subsumed or the dimensionality reduction that they offer replaced via statistical proxies.

Reflecting on whether or not a lumped model is suitable and whether or not a distributed model should be pursued instead, the results thus far suggest that a lumped model offers highly satisfactory performance. What would be gained by using a distributed model might offset that which is lost, specifically the ease of this lumped model's use and the structure of the approach being pursued for later work in attempting to create a model that generalises across geographies. Thus, a lumped model will continue to be the modus operandi of this body of work.

Following on from the work here, a range of different machine learning models and their suitability to the problem will be examined; the experimental setup along with the feature engineering will be retained as, with perhaps more appropriate model selection and tuning, improved model performance may result in a lumped model that offers prediction capability far in excess of the satisfactory whilst being easily combined with CMIP6 data.

Chapter 5

A Model Comparison

5.1 Enabling Comparisons of Extremes

With the domain appropriately defined in the previous chapter and a modelling approach established, we can now turn to increasing the model complexity and investigating other types of machine learning model, provided their application is justifiable and has merit, in order to determine the most suitable machine learning approach to the problem.

First, however, our performance metrics require some review. Performance around extremes has been lacking, such as those extreme flow events for the River Exe at Pixton in the year 2012. The NSE on its own does provide a good measure of overall fit that is sensitive to errors around those extreme events (Gupta et al., 2009). As seen with the Exe at Pixton, the relative poor performance around the higher flow extremes for the models tested in Chapter 4 distorts the NSE substantially. However, it would perhaps be useful to be better able to assess the performance around extremes through metrics that quantify the performance around those extremes of interest specifically.

Thus far, the Mean Squared Error (MSE) has been purposefully avoided, given that it would prevent the direct comparison between different catchments where the target streamflow might be orders of magnitude apart (Besaw et al., 2010). Metrics that are specifically tailored to extremes for regression problems are somewhat lacking but a weighted form of the MSE, or Squared Error Relevance Area (SERA), for imbalanced datasets, or where the predictive focus is on extreme values and rare cases, has been proposed by Ribeiro and Moniz that weights the MSE accordingly, using a relevance function (Ribeiro and Moniz, 2020). The relevance function, γ, maps the target variable domain, $\mathscr{Y}$, onto a specified scale of relevance $\gamma : \mathscr{Y} \to [0,1]$; the relevance function specified for SERA is a cubic spline function with control points defined according to statistical thresholds.

We have already established that the NSE is a normalisation of the MSE, so a new metric will be proposed that adapts elements of the automatic version of SERA and is therefore a relevance area tailoring of the NSE, which throughout this thesis will be referred to as NSE-RA. A pared back version of the automatic weighting framework will be retained; using automatic statistical thresholds at specified percentiles to define a discontinuous step function for mapping relevance that minimises the effect of the low flow events. This function is compared graphically against the normalised streamflow histogram from the River Severn at Haw Bridge catchment in Figure 5.1 and outlined formally in Equation 5.1, where the percentiles of the target distribution are denoted by ρ_y.

$$\gamma(\rho_y) = \begin{cases} 0.0 & \rho_y \leq 0.10 \\ 0.2 & 0.10 < \rho_y \leq 0.25 \\ 0.4 & 0.25 < \rho_y \leq 0.50 \\ 0.6 & 0.50 < \rho_y \leq 0.75 \\ 0.8 & 0.75 < \rho_y \leq 0.90 \\ 1.0 & 0.90 < \rho_y \end{cases} \tag{5.1}$$

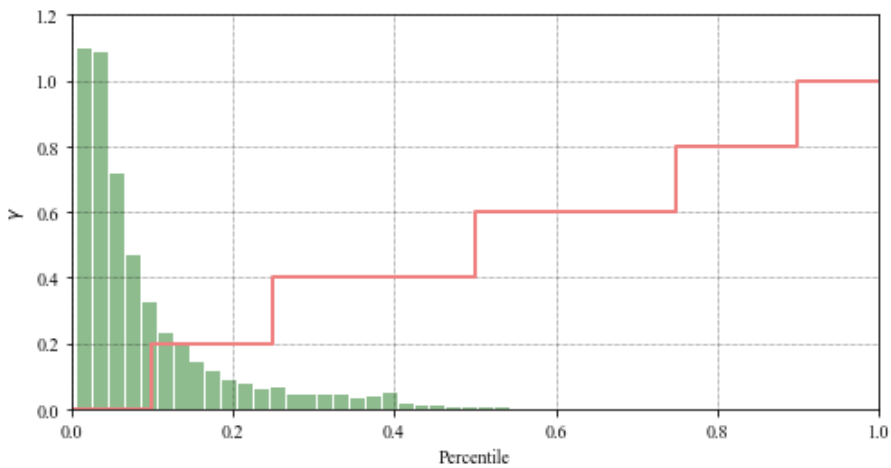


Fig. 5.1 Relevance step function for NSE-RA shown against the histogram of rescaled streamflow data for the River Severn at Haw Bridge

When used for the final results from Chapter 4, the antecedent proxy based MLP, the corresponding NSE-RA values for the Severn at Haw Bridge, the Thames at Kingston, the Avon at Bathford, and the Exe at Pixton are 0.844, 0.780, 0.827, and 0.307, respectively. With the poor performance around extremes for the Exe at Pixton, compared to baseline performance, the NSE-RA value is significantly lower than the NSE value, which was 0.709. For the other three catchments, where predictions around high extremes more closely matches observations, the NSE-RA value approaches that of the NSE.

Naturally, the NSE-RA value should likely be lower than the NSE value, as extremes, being outside of the bulk of the data, are harder to learn and fit the model to. Thus, when using NSE-RA, the objective should be to minimise the difference $NSE - NSERA$ whilst also maximising both individually.

5.2 Machine Learning Approaches

In this section, a range of different machine leaning approaches are presented, spanning the development of relatively simple neural networks into more advanced frameworks, with their application and comparative assessment for the purposes of hydrological modelling.

5.2.1 The Multi-Layer Perceptron Continued

Before expanding to more complex model types, the MLP will be revisited. The architecture of the MLP, as used in Chapter 4, was selected based on similar structures employed in reviewed literature in terms of nodes and layers; ergo, focus will shift to developing a strategy for selecting a more appropriate architecture.

The structural design of the MLP, in terms of the size of the network, must itself be optimised; after all, it would be unreasonable to expect a neural network comprised of a single neuron of being capable of expressing the same model complexity and representing the same functions as one comprised of an infinite number of neurons. However, whilst optimisation of network structure is required it is also difficult and not necessarily intuitive (Foody and Arora, 1997; Hunter et al., 2012; Mhaskar et al., 2017; Poggio et al., 2017; Schindler et al., 2016; Sildir et al., 2020) with some application in some contexts showing that shallower networks are more suitable for small datasets or statistically indistinguishable from deeper networks whilst in other applications deeper networks offer superior performance and can help avoid the curse of dimensionality, when high dimensional data can hinder model performance.

Let us assume, then, that there is an optimal number, l & n, of layers and nodes for this specific application and that number is unknown but within the following set: $[l, n \in \mathbb{N} | 1 <$

$l, n < \infty]$. Obviously, this represents an infinite number of permutations of networks and there is no analytical function that we have that could enable prediction of model performance against these two hyperparameters. Instead, a constructive method is presented, where additional layers are added into the network with the number of nodes being increased automatically with the number of nodes, n_i, for the ith layer in a network with l layers and for a set number of nodes in the first layer, n_0, is given by Equation 5.2:

$$n_i = \lceil n_0^{\frac{l-i}{l}} \rceil \tag{5.2}$$

So, for a network with 64 nodes in the first hidden layer and 3 hidden layers, the layer sizes would be $64 \rightarrow 16 \rightarrow 4 \rightarrow 1$, to the output connection of dimension 1, for a total of 84 hidden nodes.

Instead of comparing multiple different MLP architectures, according to size, with the other machine learning methods, a quick study is presented to determine whether or not varying the size of the network delivers realisable gains or losses in performance. Following the above construction method for differently sized networks, with 64 and 128 nodes in the first hidden layer, the performance for varying size is shown in Figure 5.2 for NSE and NSE-RA for the River Severn.

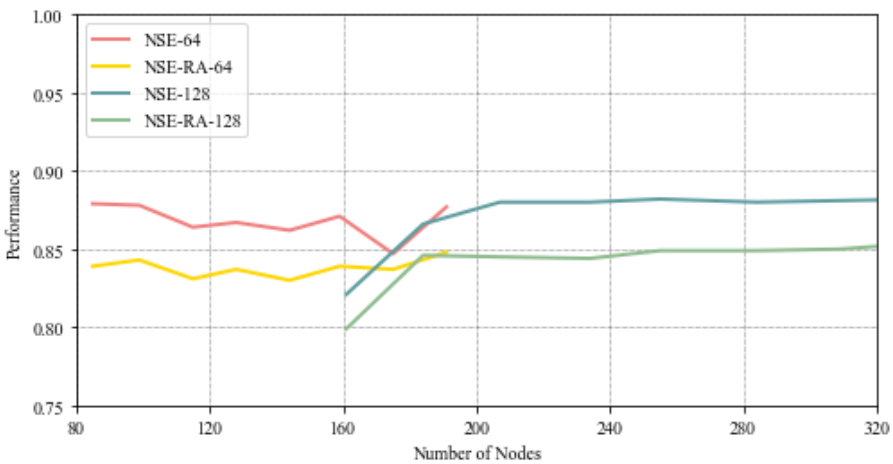


Fig. 5.2 Effect of increasing number of nodes on model performance for an MLP

Increasing the number of layers and nodes has a negligible effect on the performance of the network beyond a certain point, wherein almost doubling the network size from circa 180 nodes results in no statistically significant increase in model accuracy.

5.2.2 Recurrent Neural Networks

A Recurrent Neural Network (RNN) is an evolution of the MLP for better application to sequence modelling, where the output from previous time steps are used as additional input for processing the value of the output at the next time step, passed forward as a hidden state, whilst network weights are retained. This is expressed for a simple RNN in Equation 5.3, where the state from the previous time step is denoted $\mathbf{h}^{<t-1>}$, the current time step $\mathbf{h}^{<t>}$, with the input $\mathbf{x}^{<t>}$ and the output $y^{<t>}$. The parameters Φ_h, Φ_x, and Φ_y are separate and applied to either the input hidden state, input variables, and output hidden state along with the biases, βx and βy, prior to the use of an activation function, α, in a similar way to the standard MLP. The structure of an RNN is shown in Figure 5.3.

$$\begin{aligned} \mathbf{h}^{<t>} &= \alpha(\Phi_h \cdot \mathbf{h}^{<t-1>} + \Phi_x \cdot \mathbf{x}^{<t>} + \beta_x) \\ \& \quad y^{<t>} &= \alpha(\Phi_y \cdot \mathbf{h}^{<t>} + \beta_y) \end{aligned} \tag{5.3}$$

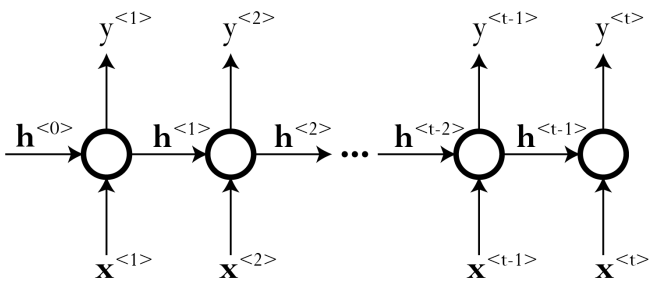


Fig. 5.3 Graphical representation of a Recurrent Neural Network of arbitrary sequence length

In spite of the advantages that the RNN has for sequence modelling, the basic RNN can struggle with exploding or vanishing gradients during optimisation and with using historical information, particularly as sequence length increases. In order to counter these deficiencies,

the Long Short-Term Memory (LSTM) unit was developed along with the Gated Recurrent Unit (GRU), a special case of the LSTM (Cho et al., 2014; Hochreiter and Schmidhuber, 1997; Kolen and Kremer, 2001). Both of these models use update gates, akin to a form of transistor circuit, to modify or retain historical information as the sequence is processed. Rather than investigating both, the GRU will be presented, due to its simpler form in having two update gates, whereas the LSTM has three.

The gates, Γ_r and Γ_u, within a GRU behave like logic gates within a circuit, acting to create an internal state, $\tilde{h}^{<t>}$, based on the incoming, $x^{<t>}$, and previous signals, $h^{<t-1>}$ and $a^{<t-1>}$, to determine the new hidden states, $h^{<t>}$ and $a^{<t>}$. Effectively, $a^{<t-1>}$ does not have a separate value but denotes a hidden state to which the activation function is applied prior to a decision gate on the extent to which this activated information is retained. The characterisation is expressed in the following Equation set 5.4, with graphical representation in Figure 5.4.

$$\begin{aligned} \tilde{h}^{<t>} &= \alpha\Big(\Phi_h\{\Gamma_r \cdot a^{<t-1>}, \mathbf{x}^{<t>}\} + \beta_h\Big) \\ \& \quad h^{<t>} &= \Gamma_u \cdot \tilde{h}^{<t>} + (1-\Gamma_u) \cdot h^{<t-1>} \\ \& \quad a^{<t>} &= h^{<t>} \end{aligned} \tag{5.4}$$

Although a range of structures for the RNN and GRU were explored, using combinations of 2, 3, and 4 hidden layers with 16, 32, 64, and 128 hidden units, only results for the highest performing model structure are presented here for simplicity of analysis and comparison with other architectures. More specifically, the results in the comparison were generated from the model with 2 hidden layers and 128 hidden units for a 14 day sequence length, as this was the easiest to train and suffered from underfitting and overfitting the least. Further discussion on the issues with these models is provided after the comparison itself.

5.2.3 Convolutional Neural Networks

The Convolutional Neural Network (CNN) is an extension to the MLP, in that the MLP is a special case of the CNN, and uses sequential convolutions over the input, whilst preserving the spatial distribution, to learn feature mappings and eventually perform classification or regression (LeCun et al., 1989).

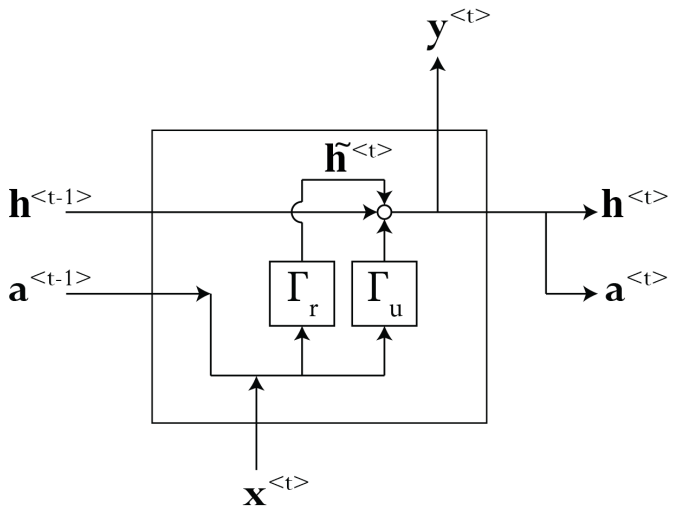


Fig. 5.4 Graphical representation of an update cell within a GRU RNN

Within each convolutional layer, for the gridded input tensor with height, d_h, width, d_w, and channels or depth, d_c, a kernel filter of size $d_n \times d_n \times d_c$ is convolved over the input to generate an output tensor of depth 1, whereby the use of multiple kernel filters, K, generates multiple output tensors that are then stacked together. So if there are k filters across l layers, with consequential convolution for a given filter defined as $y = K_L \odot \cdots \odot K_1(X)$, then a full set of convolutional layers is defined as in Equation 5.5.

$$Y = \bigodot_{i=1,\ldots,l}^{j=1,\ldots,k} K_i^j(X) \tag{5.5}$$

Typically, the sequential convolutional layers reduce the dimensionality in h and w but use increasing numbers of kernel filters to increase c. The final feature maps are then reshaped for ingress into a relatively simple MLP, as is shown in Figure 5.5. After the convolution operation within a layer, activations are applied to introduce the nonlinearity in much the same way as in an MLP between layers and additional operations, such as resampling or normalisation, can be applied to adjust the size of the tensors or counteract gradient effects.

Our rationale for applying a CNN is that the spatial structure of the data preserves the temporal aspects of the data, given that one of the spatial dimensions of the input tensor

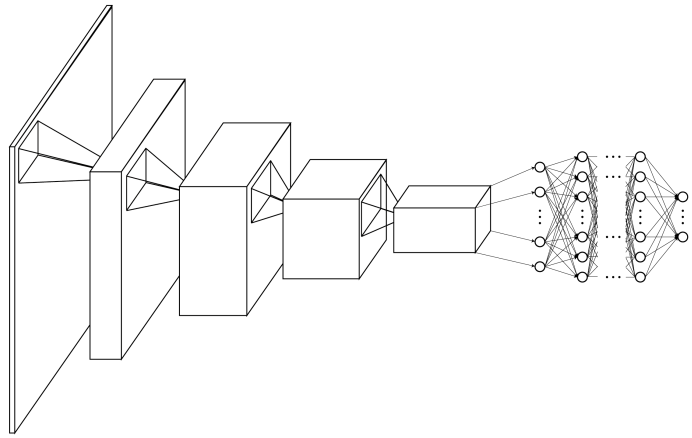


Fig. 5.5 Graphical representation of a Convolutional Neural Network of arbitrary depth and kernel size

represents time; furthermore, given the convolutional nature of the operators within a CNN, this model is close in proximity to the GLM applied in Chapter 4.

In this application of CNNs, the number of convolutional layers was varied between 2 and 5, with the number of kernels in a given layer being increased from 8 up to 256. As was the case with increasing the depth of the MLP beyond a certain point, no observable improvement in model performance was obtained by increasing the number of kernels beyond 32. The final architecture used for the comparison utilised three convolution layers with the number of filters increasing from $8 \rightarrow 16 \rightarrow 32$ before the fully connected layer; normalisation was also utilised.

5.2.4 Temporal Convolutional Neural Networks

The Temporal Convolution Neural Network (TCNN) further extends the CNN to a model capable of preserving the spatial structure of data, as with a standard CNN, whilst enabling the model to capture high-level temporal relationships, as was required in the classification of video data for which the model was developed (Lea et al., 2016). A video stream has a 3D tensor at each time step, being an image with height, width, and channels; the problem at hand in this Chapter has a 1D tensor at each time step, being the lumped climatic variables at the catchments centroid.

The TCNN captures temporal relationships using dilations; through an exponential relationship for the growth between the skipped tensors across layers over the receptive field, a TCNN is able to encode all temporal information using relatively few layers. For example,

with a dilation of 2, the receptive field at each layer doubles s.t. a full receptive field up to 64 time steps could be encoded within a single output tensor after 6 layers, with the last four layers as shown in Figure 5.6.

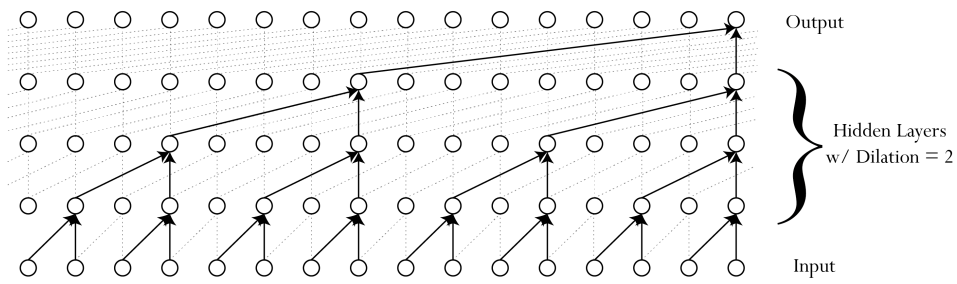


Fig. 5.6 Graphical representation of a Temporal Convolutional Neural Network with dilation=2

In terms of model architecture, the time dilation was kept equal to 2 and, keeping the same sequence length as for all other models, the number of layers required was to 4 to obtain an encoding with a receptive field greater than or equal to the sequence length. Given that 4 layers were used, a similar strategy for specifying the number of kernels was used as for the ordinary CNN, increasing as follows: $8 \rightarrow 16 \rightarrow 32 \rightarrow 64$.

5.2.5 Gaussian Processes

The models before present, in order, an evolution of the MLP. Attention now, however, turns towards a class of probabilistic models where encapsulated prior knowledge can enable better model fitting, that of Gaussian Processes (Rasmussen, 2004). A Gaussian Process (GP), Y, is defined solely by its mean, $\mu(x)$, and covariance function, $K(x,x^*)$, as characterised in Equation 5.6.

$$Y \sim \mathscr{GP}(\mu(x), k(x,x^*)) \tag{5.6}$$

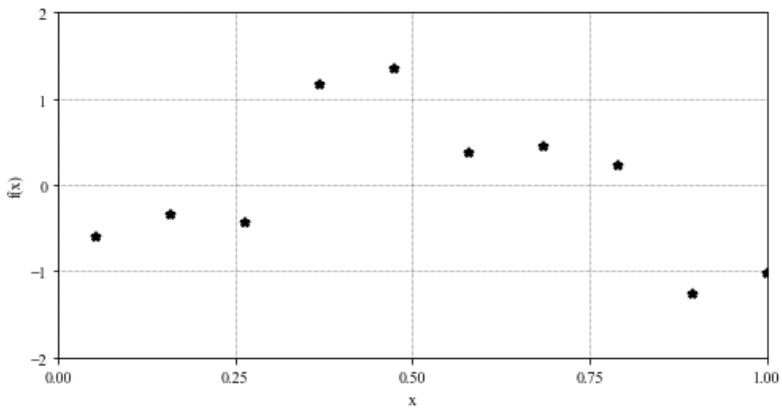
In practice, the mean function is often left equal to 0, meaning that a GP relies solely on the covariance function that essentially determines how likely a function is to take a value, with regards to the nearest known point, i.e. the given data, and is typically characterised

by its lengthscale, λ, and variance, ν. Lengthscale adjusts how rapidly the functions change in the x domain, akin to frequency, whereas variance adjusts the scaling in the target domain, more akin to amplitude. One of the most commonly used covariance functions, the Squared Exponential (SE), is expressed in Equation 5.7, from which it can be seen how these parameters adjust the function.

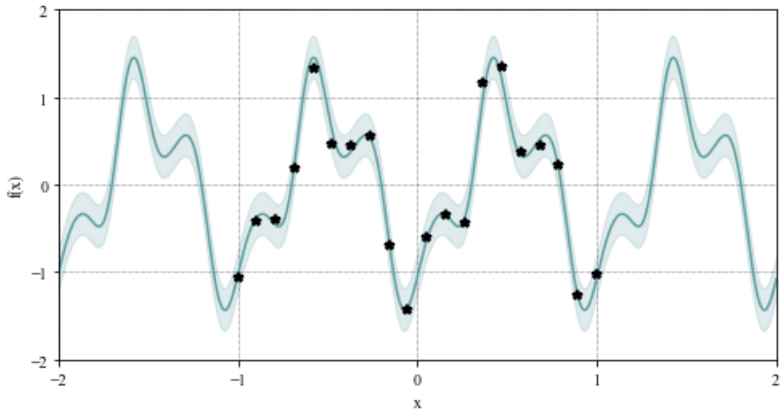
$$k(x,x^*) = \nu^2 e^{-\frac{(x-x^*)^2}{2\lambda^2}} \tag{5.7}$$

The predictive performance of a Gaussian Process, in terms of how well the model fits the data, is almost entirely determined by the covariance, or kernel, function. It is through the kernel function that the prior beliefs about which functions are likely, or the system's behaviour, are expressed. For example, consider an arbitrary periodic function with no mean trend; the most suitable kernel to select from the standard primitives would be a periodic kernel. An example of a set of data points sampled from a periodic function along with the corresponding GP fit using a periodic kernel is shown in Figure 5.7.

The fit for the Basic Linear Model in Chapter 4 was poor but it did reveal something about streamflow response to precipitation: there is a strong linear component in that as precipitation increases, so too does streamflow by a similar relative amount. Thus, one can conclude that a linear kernel is a sensible starting point. Furthermore, use of a linear kernel, in contrast with the standard SE or Matérn kernels, allows for non-local structure to be learned and, when combined with an SE or Matérn kernel allow for more complex, expressive structure (Duvenaud et al., 2011). For the more extreme streamflow events, said response is more extreme than perhaps a simple linear response, so an additional component is warranted that allows for higher order growth. As for enabling local and inceasing variation, the Matérn kernel family are useful in geostatistical work with the exponential being a common correlation function and the roughness of the Matérn kernel being more suitable and resembling physical problems than the overly smooth SE kernel (Guttorp and Gneiting, 2006). Thus the kernel defined for this problem is an additive kernel, using linear, k_{lin}, and Matérn, k_{Mat}, kernels, with structure outlined in Equation set 5.8 applied to all input variables with the lengthscale, λ, and variance, ν, optimised per variable.



(a) Sample data from a periodic function with added Gaussian noise



(b) A GP model fitted to the same data using a periodic kernel

Fig. 5.7 Gaussian Process model fit for a periodic kernel based on the prior knowledge that the underlying data is generated by a periodic function

$$
\begin{aligned}
k_{additive} &= (k_{Mat} \times k_{lin}) + (k_{Mat} \times k_{lin} \times k_{lin}) \\
\& \quad k_{Mat}(x,x^*) &= \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\frac{\sqrt{2\nu(x,x^*)}}{\lambda} \right)^{\nu} \cdot K_{\nu} \cdot \left(\frac{\sqrt{2\nu(x,x^*)}}{\lambda} \right) \\
\& \quad k_{lin}(x,x^*) &= \nu^2 (x-c)(x^*-c)
\end{aligned}
\tag{5.8}
$$

Where for k_{Mat}, the Matérn kernel (Matérn, 1986), K_{ν} is the modified Bessel function of the second kind and for k_{lin}, the linear kernel, c represents the linear constant. Samples drawn from the kernel function in one dimension are shown in Figure 5.8, highlighting the nonstationarity of the kernel and how the function $f(x)$ evolves as x^* moves away from the starting point $x = 0$.

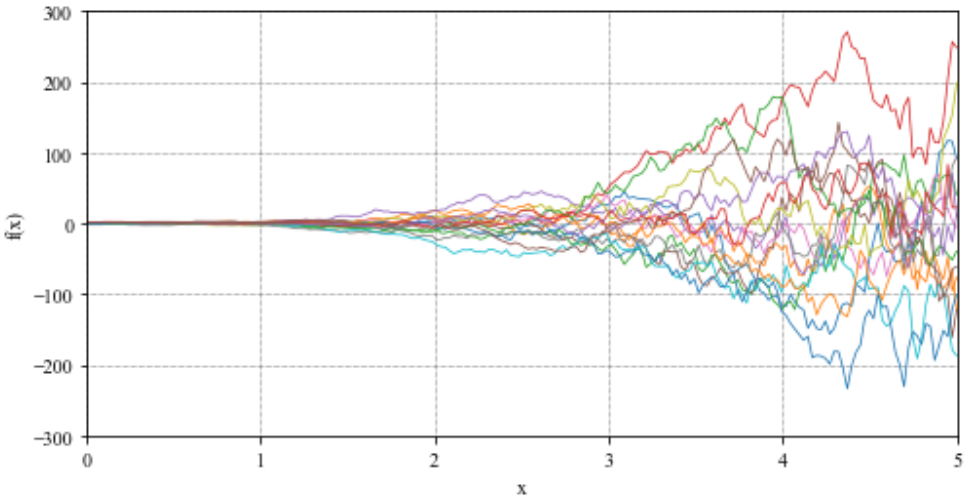


Fig. 5.8 Function samples drawn from the additive kernel specified for the problem of predicting streamflow, showing how the functions diverge as the value of x^* moves away from x and continue to do so under the influence of the linear part of the kernel

Fitting a Gaussian Process is done via Maximum Likelihood Estimation of the hyperparameters, as shown in Equation 5.9, with the set of optimised hyperparameters, Θ', then used along with the given data to make further predictions at the matrix of test points, $\mathbf{X}_*$, as shown in Equation 5.10 (Rasmussen, 2004).

$$\Theta' = \underset{\Theta}{argmax}\ log\ P(\mathbf{y}|\mathbf{X},\Theta) \tag{5.9}$$

$$P(f_*|\mathbf{y},\mathbf{X}_*,\mathbf{X},\Theta') = \mathscr{N}\left(f_*|\mathbf{K}_*(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y},\mathbf{K}_{**}-\mathbf{K}_*(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{K}_*^T\right) \tag{5.10}$$

Where $\mathbf{K}$ is the matrix of all pairwise covariance function evaluations of observations in the training set, $\mathbf{K}_*$ is that for the training and test set, and $\mathbf{K}_{**}$ is that for the test set, σ is the standard deviation, $\mathbf{I}$ is the identity matrix, $\mathbf{y}$ are the test observations, and f_* is the function drawn. Evaluating the log likelihood in Equation 5.12 (Rasmussen, 2004) requires taking a Cholesky decomposition, $\mathbf{L}$, as in Equation 5.11.

$$\begin{aligned} \mathbf{L} &= chol\big(\mathbf{K}+\sigma^2\mathbf{I}\big) \\ s.t.\ \ \mathbf{L}\mathbf{L}^T &= \mathbf{K}+\sigma^2\mathbf{I} \end{aligned} \tag{5.11}$$

$$\begin{aligned} log\ P(\mathbf{y}|\mathbf{X},\Theta) &= log\ \mathscr{N}(\mathbf{y}|0,\mathbf{K}+\sigma^2\mathbf{I}) \\ &= -\frac{n}{2}n\ log\ 2\pi - \sum_i log\ \mathbf{L}\mathbf{L}_{ii} - \frac{1}{2}||\mathbf{L}^{-1}\mathbf{y}||^2 \end{aligned} \tag{5.12}$$

For n observations in the training set, the order of computational complexity for all pairwise evaluations in the covariance matrix is $\mathscr{O}(n^2)$ and for the Cholesky decomposition of a matrix $\mathscr{O}(n^3)$. Therefore, the overall order computational complexity is $\mathscr{O}(n^3)$.

5.2.6 Sparse Variational Gaussian Processes

The major drawback of a GP model is the high computational complexity with regards to optimisation (Liu et al., 2019a); for applications with few data points this issue is largely negligible but an ordinary GP will quickly become intractable as we increase the number of data points n from the 1×10^5 range for single catchments to 1×10^7 for a generalising model, discussed further in Chapter 6, which would increase the computational demand by 6 orders of magnitude.

In order to improve the model such that it is more tractable, one might consider whether or not every pairwise evaluation in the covariance matrix is necessary and the same for every evaluation in Cholesky decomposition; in other words, if we can either introduce sparsity or approximations into the model, then the computational complexity is reduced. Sparse Variational Gaussian Processes (SVGP) are a modification to the GP framework that reduces the computational complexity for n data points from $\mathscr{O}(n^3)$ to $\mathscr{O}(n \cdot m^2)$ using m pseudo-data points (Hensman et al., 2013; Titsias, 2009).

Consider a GP with observations $\{\mathbf{x}_i\}_{i=1}^n$ where we introduce pseudo observations $\{\mathbf{z}_i\}_{i=1}^m$, where $m < n$ and all possible pairs made from combined set of all $\mathbf{x}$ and $\mathbf{z}$ are, as per an ordinary GP, considered to have some dependency through the covariance function.

$$P(\mathbf{x},\mathbf{z}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix};\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}\mathbf{K}_{xx} & \mathbf{K}_{xz}\\ \mathbf{K}_{zx} & \mathbf{K}_{zz}\end{bmatrix};\right) \tag{5.13}$$

Then, removing the dependency between all pairs of $\mathbf{x}$ s.t all pairs of $\mathbf{x}$ are conditionally independent from one another given the set $\{\mathbf{z}_i\}_{i=1}^{m}$, or $\mathbf{K}_{xx}|\mathbf{z} = 0$, implying that this set of pseudo observations encapsulates all of the correlations and relationships between the original set of observations.

In order to calibrate the model to ensure that the approximate, sparse form of the GP matches that of the original distribution, the difference between the distribution of the original and the sparse form must be minimised. Thus, the Kullback–Leibler Divergence, being a measure of how different a pair of probability distributions are, will be used and is expressed in Equation 5.14 (Kullback and Leibler, 1951), which we can frame in terms of the observed and predicted probability distributions, U and V respectively, of the random variable, x.

$$D_{KL}(U||V) = \sum_{x} U(x)\log\left(\frac{U(x)}{V(x)}\right) \tag{5.14}$$

We note that whilst the KLD is a measure of how similar or dissimilar one probability distribution is from another, it is not a distance metric and, thus, not symmetric. Rewriting for the GP model and the KLD to be minimised is as follows, in Equation 5.15.

$$\underset{V(\mathbf{z}),\left\{V(\mathbf{x_i}|\mathbf{z})\right\}_{i=1}^{n}}{argmin} \quad D_{KL}\big(P(\mathbf{x},\mathbf{z})||V(\mathbf{z})\prod_{i=1}^{n}V(\mathbf{x_i}|\mathbf{z})\big) \tag{5.15}$$

Therefore, with the approximate model optimised in terms of its representation of the original underlying model to arrive at the Equations 5.16 & 5.17, the process for solving a Gaussian Process can be followed for Equation 5.18.

$$V(\mathbf{z}) = \mathcal{N}(\mathbf{z};0,\mathbf{K}_{zz}) \tag{5.16}$$

$$V(\mathbf{x_i}|\mathbf{z}) = \mathscr{N}\left(\mathbf{x_i}; \mathbf{K}_{x_i z}\mathbf{K}_{zz}^{-1}\mathbf{z}, \mathbf{K}_{x_i x_i} - \mathbf{K}_{x_i z}\mathbf{K}_{zz}^{-1}\mathbf{K}_{z x_i}\right) \tag{5.17}$$

$$P(\mathbf{y_i}|\Theta) = \mathscr{N}\left(\mathbf{y}; 0, \mathbf{K}_{xz}\mathbf{K}_{zz}^{-1}\mathbf{K}_{zx} + D + \sigma_y^2 I\right) \tag{5.18}$$

Given $D_{ii} = \mathbf{K}_{x_i x_i} - \mathbf{K}_{x_i z}\mathbf{K}_{zz}^{-1}\mathbf{K}_{z x_i}$. Thus, the only matrix inversion is for $\mathbf{K}_{zz}^{-1}$, hence the lowered computational complexity; however, the model becomes parametric on the pseudo points and less a Gaussian Process, so the Variational Free Energy method was proposed. With the true posterior being intractable, a lower bound, $\mathscr{F}(\Theta)$, is introduced on the likelihood, $log\ P(\mathbf{y}|\Theta)$, via Jensen's inequality (Hensman et al., 2015; Titsias, 2009), as expressed in Equation 5.19.

$$\begin{aligned}\mathscr{F}(\Theta) &= \int df\, V(f)\, log\, \frac{P(f|\mathbf{y},\Theta)P(\mathbf{y}|\Theta)}{V(f)} \\ &= log\, P(\mathbf{y}|\Theta) - D_{KL}\big(V(f)||P(f|\mathbf{y})\big)\end{aligned} \tag{5.19}$$

If an approximate form of the posterior for $V(f)$ is assumed, such that it is split into a finite number of pseudo-points, $\mathbf{z}$, and the remaining infinite set, $f_{\neq \mathbf{z}}$, then it can be expressed as shown in Equation 5.20 before being used to rewrite $\mathscr{F}(\Theta)$ as in Equation 5.21 before being maximised, thus reducing the D_{KL} (Hensman et al., 2015; Titsias, 2009).

$$\begin{aligned}V(f) &= V(\mathbf{z}, f_{\neq \mathbf{z}}) \\ &= V(f_{\neq \mathbf{z}}|(\mathbf{z})V((\mathbf{z}) \\ &= P(f_{\neq \mathbf{z}}|(\mathbf{z})V((\mathbf{z})\end{aligned} \tag{5.20}$$

$$\begin{aligned}\mathscr{F}(\Theta) &= \int df\, V(f)\, log\, \frac{P(\mathbf{y}|f,\Theta)P(\mathbf{z}}{V(\mathbf{z})} \\ &= \langle log\, P(\mathbf{y}|f,\Theta)\rangle_{V(f)} - D_{KL}\big(V(\mathbf{z})||P(\mathbf{z})\big)\end{aligned} \tag{5.21}$$

The advantage of this Variational Free Energy method is that the pseudo data points, in terms of their location in the input domain, are pure variational parameters and keeps the approximation and the prior beliefs about the model separate, in other words being more non-parametric.

Typically, the number of pseudo points, m, should be equal to the number of inflection points in the underlying function (Snelson and Ghahramani, 2005; Turner, 2017); given the high dimensionality of our problem, evaluating the number of inflection points is problematic. Thus, if we let $m = n^{1/2}$, then $\mathscr{O}(n \cdot m^2)$ simplifies to $\mathscr{O}(n^2)$ and offers a reasonable assumptive estimate whilst significantly decreasing the computational complexity.

5.2.7 Neural Processes

Gaussian Processes have many advantages but a key disadvantage is the aforementioned scalability problem. Furthermore, the key strength of a Gaussian Process could also be its weakness, in that one must be able to specify the kernel in order to attain strong model performance. Conditional Neural Processes (Garnelo et al., 2018) offer a hybrid, of sorts, wherein some of the strengths of both MLP and GP models are combined, achieving rapid learning from data and low computational demand, whilst still being a probabilistic model and supporting relative agnosticism with regards to the model structure.

The architecture of a NP involves 3 components, the first being an MLP encoder, f_e, which takes input and output context pairs, (x_c, y_c), and learns a representation, r_c, as per Equation 5.22:

$$r_c = f_e(x_c, y_c) \tag{5.22}$$

This encoding is then aggregated and parameterised to obtain a normal distribution of the parameter z, as in Equation 5.23:

$$P(z|x_i, y_i) = \mathcal{N}\left(\mu_i(r), \sigma_i^2(r)\right) \tag{5.23}$$

Before z is sampled and used as input alongside the target, x_t, to a decoder MLP, f_d, to obtain an output sample from the predictive distribution of y_t, Equation 5.24:

$$y_t' = f_d(x_t, z) \tag{5.24}$$

The resulting model is one that is scalable and a conditional distribution over functions. As such, optimisation is performed through minimising the combined negative log likelihood (NLL), as expressed in Equation 5.25, and KLD, as expressed in Equation 5.14.

$$NLL = -\sum_{i=1}^{n} \left\{ y_i \cdot ln\left(P(y_i')\right) + \left(1 - y_i\right) \cdot ln\left(1 - P(y_i')\right) \right\} \tag{5.25}$$

Where the NLL is presented terms of the observation y_i and predictand y_i'. In all, the combined NLL and KLD are suitable for this form of optimisation, wherein our model performance and predictive capability improves as the predicted values and distribution approach that of the observed; this model therefore gives a conditional distribution over functions and enables the prediction of the most likely underlying function with the variation treatable as uncertainty about that function. Finally, the computational complexity for the Neural Process is given as $\mathcal{O}(n+m)$ for n data points in the context set and m data points in the target set, a realisable improvement over the computational complexity of Gaussian Processes.

5.3 The Comparison

Each of the machine learning models detailed above were applied using the same training and test data split for each of the catchments used in Chapter 4 and with the same feature space, albeit modified in terms of data ingress to suit the specific method, such as structuring the inputs appropriately for the sequence models. The results for each of the four catchments individually are shown in Tables 5.1, 5.2, 5.3, and 5.4, whilst an average performance for

each of the models across these four catchments is shown in Table 5.5. Also included in Table 5.5 is the average training set fit, highlighting the tendency of a model to overfit in spite of regularisation or other methods to prevent overfitting. Furthermore, a subset of the results, specifically the test set year 2012 for the River Severn at Haw Bridge, is depicted graphically in Figure 5.9.

Table 5.1 Model prediction performance comparison for the River Severn at Haw Bridge

Catchment	Metric				
	NSE	NSE-RA	Flow Quantile $(m^3 s^{-1})$		
			Q_{25}	Q_{50}	Q_{75}
Observations	-	-	36.8	64.8	134
Shallow MLP	0.875	0.844	34.8	66.7	135
Deep MLP	0.878	0.850	33.7	68.0	132
RNN	0.851	0.823	36.3	72.1	136
GRU	0.829	0.789	33.7	64.3	130
CNN	0.881	0.836	35.9	67.7	127
TCNN	0.786	0.577	35.3	68.9	129
NP	0.874	0.855	36.9	69.5	127
GP	0.830	0.772	34.7	68.9	135
SVGP	0.854	0.786	39.1	82.6	156

Table 5.2 Model prediction performance comparison for the River Thames at Kingston

Catchment	Metric				
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})		
			Q_{25}	Q_{50}	Q_{75}
Observations	-	-	11.7	29.9	79.5
Shallow MLP	0.853	0.780	12.4	35.9	86.7
Deep MLP	0.867	0.805	12.8	35.4	85.2
RNN	0.826	0.769	14.6	35.8	92.7
GRU	0.785	0.726	12.5	36.8	87.4
CNN	0.845	0.734	11.8	34.5	83.4
TCNN	0.797	0.629	13.7	34.8	82.5
NP	0.865	0.787	15.2	36.2	87.8
GP	0.818	0.780	13.4	38.3	85.6
SVGP	0.853	0.780	14.7	40.5	92.5

Table 5.3 Model prediction performance comparison for the River Avon at Bathford

Catchment	Metric				
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})		
			Q_{25}	Q_{50}	Q_{75}
Observations	-	-	4.24	8.53	22.8
Shallow MLP	0.874	0.827	4.38	9.68	26.4
Deep MLP	0.867	0.831	4.20	9.64	26.6
RNN	0.858	0.796	5.16	11.1	26.1
GRU	0.802	0.748	4.59	9.36	24.4
CNN	0.820	0.665	4.41	9.87	24.6
TCNN	0.830	0.731	3.91	9.30	24.2
NP	0.873	0.837	4.06	11.0	25.8
GP	0.853	0.797	3.97	10.8	25.2
SVGP	0.863	0.840	3.45	11.5	26.9

Table 5.4 Model prediction performance comparison for the River Exe at Pixton

Catchment	Metric				
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})		
			Q_{25}	Q_{50}	Q_{75}
Observations	-	-	1.26	2.30	5.44
Shallow MLP	0.709	0.307	1.39	2.80	5.47
Deep MLP	0.763	0.533	1.29	2.60	5.31
RNN	0.744	0.539	1.49	2.56	5.30
GRU	0.723	0.556	1.46	2.53	5.12
CNN	0.737	0.386	1.56	2.69	5.34
TCNN	0.666	0.192	1.33	2.78	5.24
NP	0.764	0.560	1.60	2.61	5.65
GP	0.676	0.546	1.60	2.74	5.11
SVGP	0.754	0.550	1.68	2.90	5.54

Table 5.5 Average model fitting and prediction performance comparison

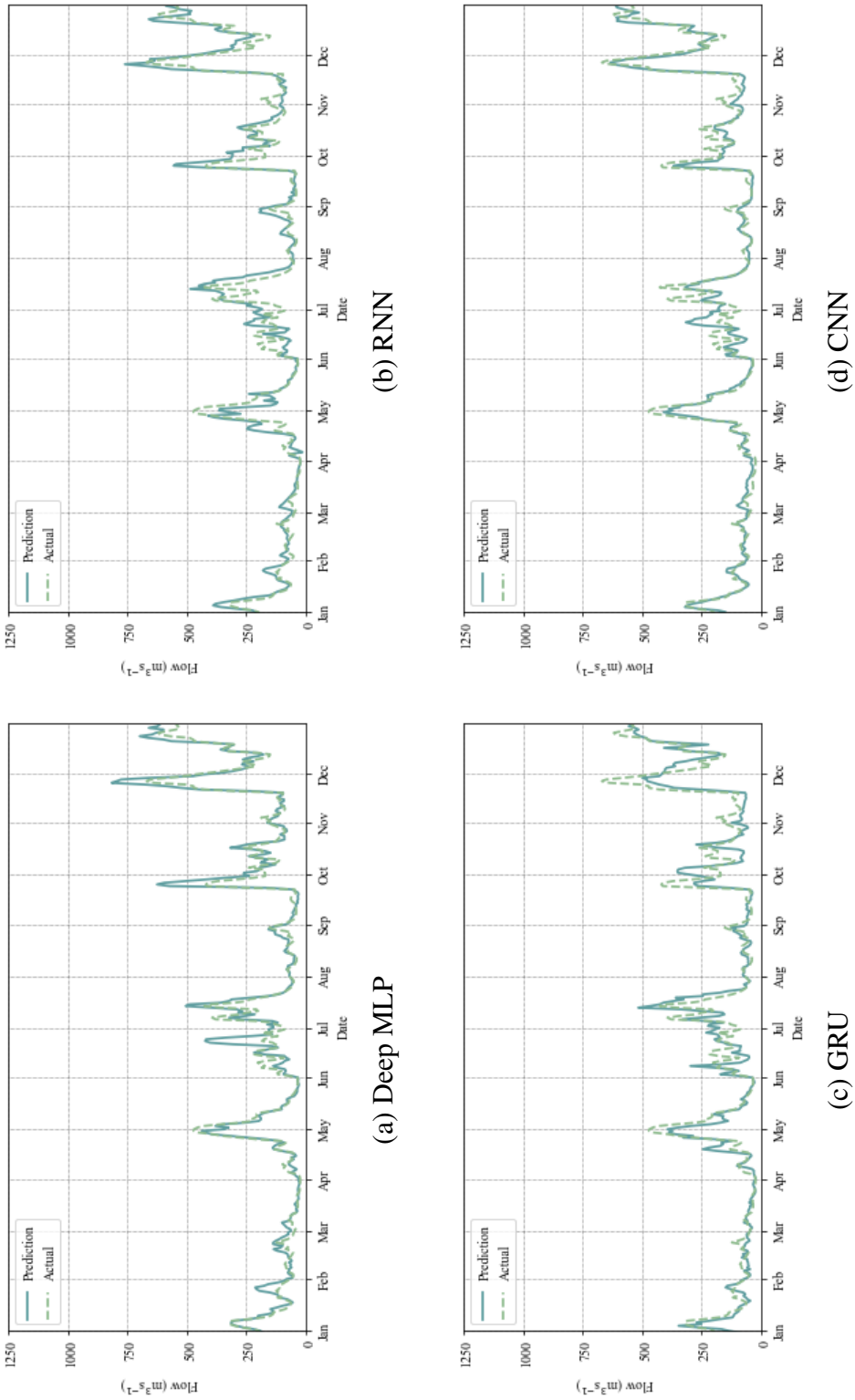
Model	Metric			
	Training Set		Test Set	
	NSE	NSE-RA	NSE	NSE-RA
Shallow MLP	0.869	0.773	0.828	0.690
Deep MLP	0.920	0.813	0.844	0.755
RNN	0.828	0.754	0.820	0.732
GRU	0.931	0.892	0.785	0.705
CNN	0.844	0.729	0.821	0.655
TCNN	0.794	0.654	0.770	0.532
NP	0.839	0.774	0.844	0.760
GP	0.804	0.735	0.794	0.724
SVGP	0.821	0.785	0.831	0.739

Standout model performance is obtained with the Deep MLP, the NP, and the SVGP. The MLP is easy to train and fit as well as specify hyperparameters and architecture; its simplicity

is its strength. The GP and SVGP models perform well because they are Bayesian and the practitioners prior beliefs about the domain can be directly and easily encoded through the kernel function; the advantage that the SVGP appears to have over the ordinary GP is due to the high computational demand of the ordinary GP limiting the amount of time available for training. Conversely, if nothing was known about the domain, then a GP would be a poor choice. Rather than recommend automatic methods for selecting kernels, our belief is that, in the absence of a priori knowledge, a Neural Process is likely the superior approach. In fact, the NP was the best all round approach with the highest average NSE and NSE-RA for its mean prediction. However, in contrast with the GP, the NP model's variance, after training is complete, is reduced significantly, towards being almost neglible. Therefore, where we might prefer that the model was less certain, such as around those high flow extremes that occur less frequently, the converse was true and this behaviour is undesirable.

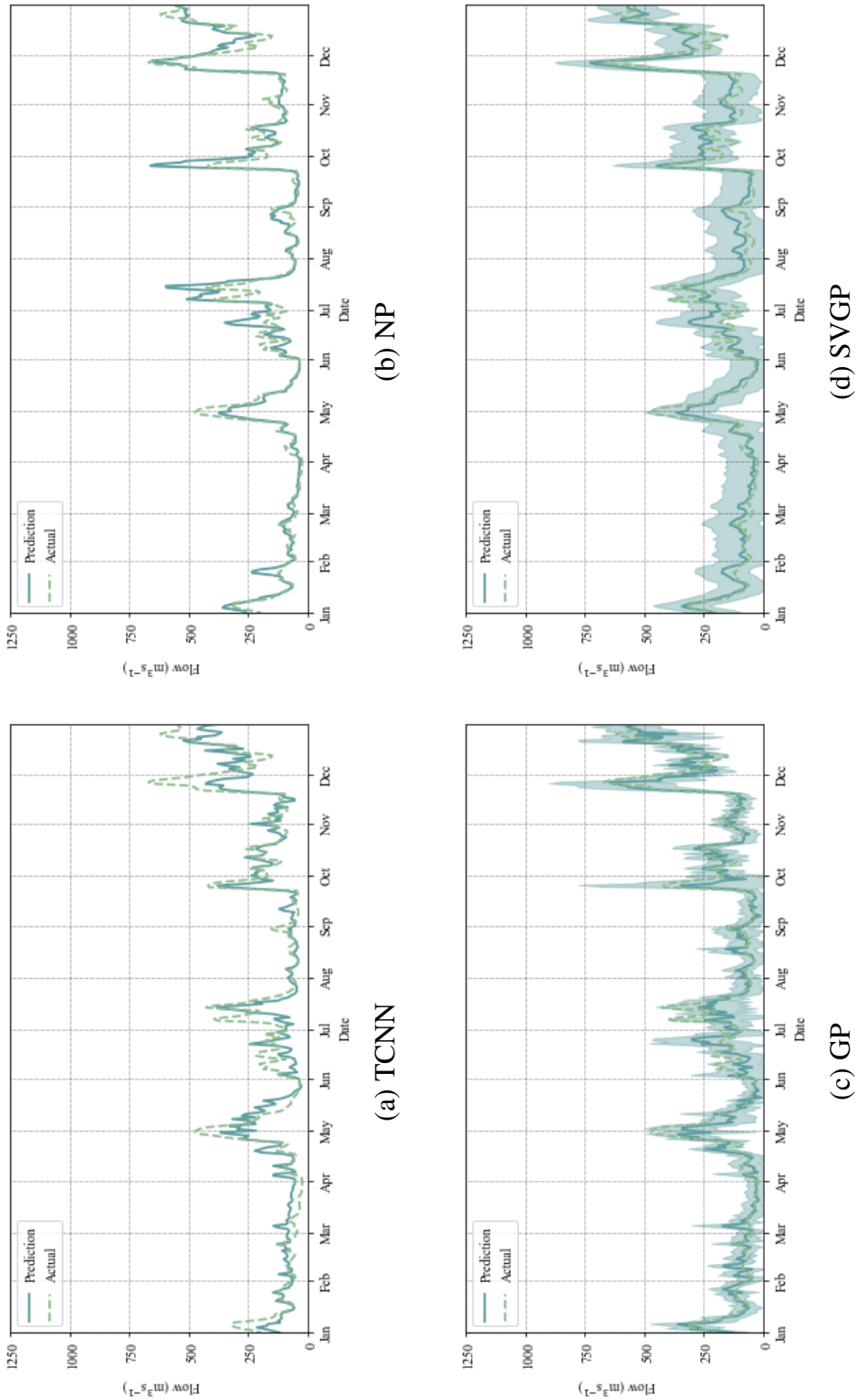
With regards to the GRU model, it has been noted that recurrent networks are susceptible to overfitting, which can be difficult to correct (Bengio et al., 1994; Pascanu et al., 2013); in this study, stopping training early, regularisation, and dropout were all investigated but to limited and the same effect, in that the model struggles to predict above a threshold in the target domain when attempting to reduce overfitting, irrespective of the method used. This thresholding effect limited the performance of the models around extremes, resulting in the comparatively weaker NSE-RA, and was evident across all of the architectures outlined earlier in this chapter but, interestingly, was not the case with the basic RNN.

The relative strength, or lack thereof, of the TCNN, was somewhat surprising, given that its structure seems to lend itself towards this application, essentially being the encoding of a full temporal field; but, as with the GRU, the TCNN model is prone to overfitting and difficult to regularise. The large number of hyperparameters and architectural layouts decreases the viability of a systematic search through these permutations to find an optimal configuration. To a lesser extent, the same issue was present with the ordinary CNN, hence the lower NSE-RA values for both of these models.



(a) Deep MLP (b) RNN (c) GRU (d) CNN

Fig. 5.9 Streamflow time series for the test set year, 2012, with predictions and observations for the River Severn at Haw Bridge using each model type



(a) TCNN (b) NP (c) GP (d) SVGP

Fig. 5.9 Streamflow time series for the test set year, 2012, with predictions and observations for the River Severn at Haw Bridge using each model type

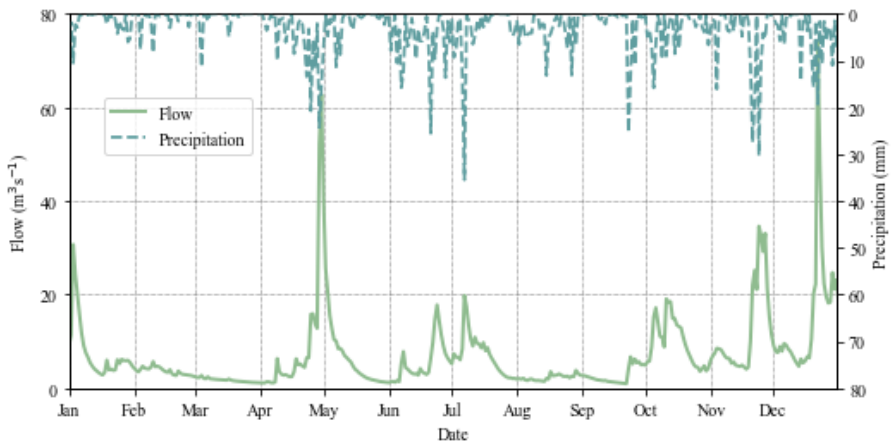
5.4 Anomalies, Refinements, & Further Work

5.4.1 Pattern Representation in Training

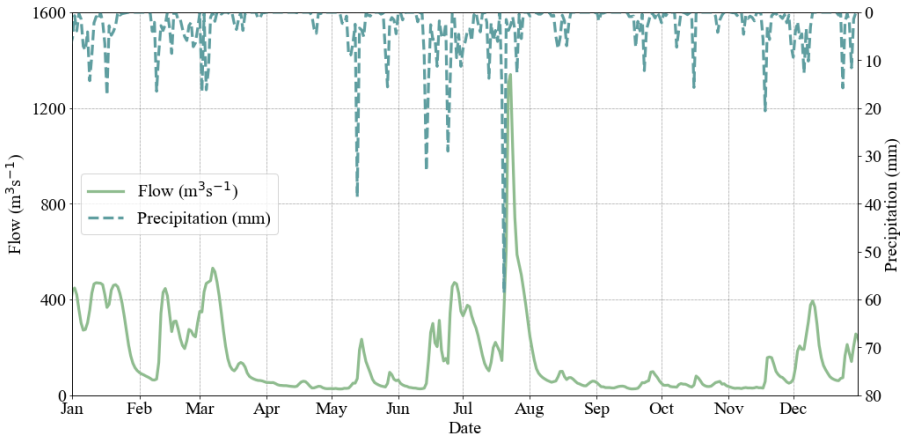
Average performance across all models for the Severn at Haw Bridge, the Thames at Kingston, and the Avon at Bathford using the NSE-RA metric are 0.792, 0.754, and 0.785, respectively, but for the River Exe at Pixton the performance is 0.463; model performance around extremes for the River Exe at Pixton is comparatively poor but, given the strength for the other three catchments, fundamental issues with the modelling approach is unlikely to be the cause.

Training set composition, in terms of how well it represents patterns and behaviours that the model is likely to be utilised for, is key (Batista et al., 2004; Foody et al., 1995; Goodfellow et al., 2016) and therefore a suitable place to investigate. Upon said further investigation, the data points that are primarily driving the poor performance for the Exe at Pixton are a few key dates in 2012; 2012 was notable for having significantly high rainfall, being one of the years with the highest rainfall on Met Office record and especially so in April and June (Parry et al., 2013). Inclusion of this year within the training set, such that the test set is now for years 2010 to 2019 less 2012, results in a significant improvement in general performance and extreme performance, with NSE improving from 0.764 to 0.780 and NSE-RA improving from 0.546 to 0.673 for the NP model. Conversely, removing 2012 from both the test and training sets, such that the model is neither learning from this year nor is it being evaluated on it, resulted in an NSE of 0.775 and NSE-RA of 0.624. Thus, the appropriate conclusion is that there is important information from the hydrological patterns of 2012 that improves the generalising capabilities of the model but not to levels of the other models, even though the remaining test set has comparatively few extremes.

The problem is that the Exe catchment was more affected by extreme rainfall than in June, when compared against previous years, than April, making June the more unusual extreme, and yet the corresponding increase in hydrological response was lower and far more accurately predicted for June. In Figure 5.10, the rainfall and streamflow are plotted together, showing this discrepancy between peak locations. For comparison, also shown in Figure 5.10 is the extreme flow event for the River Severn in July 2007, exceeding all other extreme flow events on record by at least 50%, and which corresponds to a significant increase in rainfall in the catchment.



(a) Exe at Pixton



(b) Severn at Haw Bridge

Fig. 5.10 Streamflow time series for the River Exe for the year 2012 with observed flow and catchment average daily precipitation, highlighting the anomalous peaks for the River Exe versus the rain-driven peaks for the River Severn occurring in 2007 at the time of the Tewkesbury floods (Marsh and Hannaford, 2007)

Ordinarily, one might question the veracity of the flow data recorded by the gauging station; however, as seen in Chapter 4, this phenomena extends across other gauges at different locations along the same river. No information on this catchment's specific internal mechanics could be found through the literature searched for this thesis but, given the consensus between the cumulative rainfall used in the dataset for training the model and the Met Office records, the only conclusion to draw is that the River Exe is responding to other variables beyond rainfall, and other climatic variables, in these extremes. We instead assume that there is some unknown mechanic, likely human, at play driving streamflow either under

specific conditions, such as the discharge from reservoirs or other human controlled storage systems; this, therefore, would not necessarily correlate with statistical climatic thresholds alone and may be driven primarily by unknown variables that are not being measured at all.

5.4.2 NSE-RA Refined

The NSE-RA metric developed is a useful quantifier in predicting performance around extremes in this context, in that it further increases the sensitivity of NSE to extremes but the implementation was crude and the weighting strategy would have to be reformulated for distributions with two tails or unusual cases at various points through the domain, for example a distribution arising from a mixture of Gaussians.

A potential refinement to the NSE-RA metric, and one that could be further applied to SERA, is to use an inferred probability distribution as the basis for the relevance function. If we have some arbitrary training dataset that is also representative of the expected values in situ, then the reflected probability distribution provides a function that appropriately weights the data according to how likely it is to occur. Concretely, the probability function, $U(x)$, reflected in the x axis, adjusted by a scaling factor, κ, and then added to one results in a continuous relevance function that can be used to automatically remap the error, as formulated in Equation 5.26:

$$\gamma = -\frac{U(x)}{\kappa} + 1 \tag{5.26}$$

For example, consider an arbitrary training dataset that is observed to fit a normal distribution with inferred mean μ and standard deviation σ, then from the formulation of normal distribution, as shown in Equation 5.27, then the maximum of the distribution, obtained setting $x = \mu$, provides the scaling factor, κ, for the normal distribution is obtained in Equation 5.28.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{5.27}$$

$$\kappa = \frac{1}{\sigma\sqrt{2\pi}} \tag{5.28}$$

The resulting relevance function for this dataset is shown in 5.11. Given the weighting distribution is matched to the data distribution, it should be clear how this metric is sensitive to all extremes and extends well to cover less standard data distributions, such as the aforementioned Gaussian mixtures. And with NSE being similar to R^2, one can conclude that it could be a suitable metric for all regression problems where extremes are of interest and could even be used as a loss function in regression problems to deliver superior generalisation capability.

Furthermore, this metric could serve as the cost function in a machine learning model optimised through gradient descent methods, replacing the MSE loss function to provide one that penalises performance around higher probability densities and rewarding performance where that density is lower. Unlike the previous iteration of NSE-RA, this function would be differentiable for most probability distributions but, for those that might not be, fitting a Gaussian or mixture of Gaussians could accurately approximate the underlying distribution well. Consequently, models optimised using reflective R^2 metric could realise gains in performance.

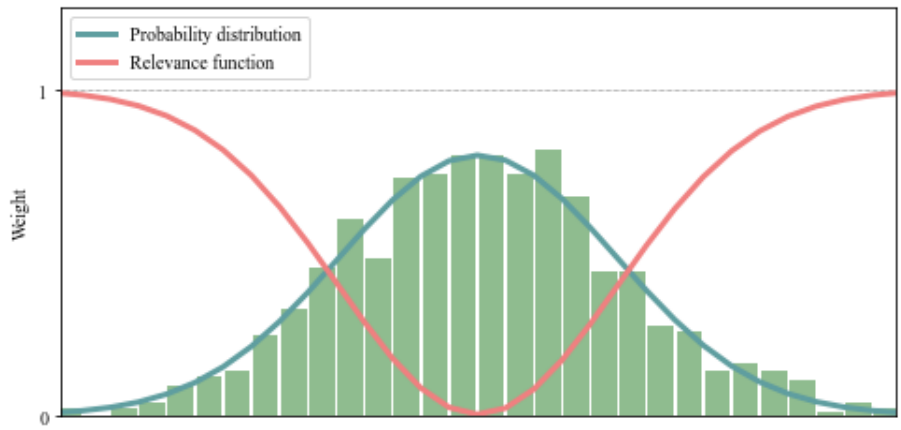


Fig. 5.11 Proposed weighting function for a refined NSE-RA, based on a probability distribution inferred from data

5.4.3 Further Work

In summary, between most models the difference in performance was relatively minor compared to the performance gains attained through the feature engineering work conducted in Chapter 4; this reinforces the view purported that quality of data is significantly more important than model complexity (Lawrence, 2019). However, better performance is obtained

when one is able to encode prior beliefs, as with the GP models, or the structure of the model is more easily justifiable. For those models that underperformed, that they might reach the same level of performance cannot be ruled out but it is a reflection of the difficulty of optimising their architecture and their training.

Optimisation for those underperformers and systematic, that is to say grid, searches through hyperparameters was a lengthy process; and, depending on the step size in that grid search, it would be difficult to assume attainment of the optimal values for those hyperparameters. This is intuitive, given that the number of permutations for n hyperparameters with m settings would be m^n and, if setting m to lower the computational demand from the total number of model evaluations requiring investigation, there is a distinct possibility that the grid search would overstep the optimum. Retrospectively, a more elegant framework for said optimisation could have been sought, with Bayesian Optimisation methods, which can include Gaussian Processes, having proven performance in selecting hyperparameters that maximise model performance in relatively few iterations (Brochu et al., 2010; Snoek et al., 2012).

Another avenue that would have been worth exploring are the additional models from the GP family. The comparative performance of the base GP and RNN suggest that a hybrid, beyond those that are investigated within Chapter 7 of this thesis, would likely be a viable approach; such models already exist in the form of Recurrent Gaussian Processes, which were demonstrated to have superior performance to standard GP models and MLPs for sequential problems (Mattos et al., 2016) in an experimental setup. Future work should explore this framework for its potential application to sequential climate problems.

5.5 Optimal Architecture

If the primary outcome of the preceding chapter was data centric, then the primary outcome of this one is model centric. A comparison between different modelling approaches has been provided and we have explored the suitability of these architectures, in terms of the implementation complexity and applicability to the task at hand.

To further the point about a data centric approach being critical to machine learning success, we showed that no model can overcome deficiencies within the training set, using the River Exe as our example. By extension, the importance of domain knowledge cannot be understated and we illustrated this through the design of a Gaussian Process kernel for delivering strong performance around both baseline and extreme predictions.

Finally, we have described a new statistical measure for the analysis of model performance when it comes to extreme values; we also proposed an improvement to this metric, making

it non-parametric and reflective of the domain probability distribution. Given that extreme values are of considerable interest in many research fields, we expect that this metric may have widespread potential.

Chapter 6

Generalising Across Catchments

The focus of this chapter is turned to the development of a model that can learn from geographies with sufficient historical records of streamflow and predict for those without that data. As discussed in the introduction, this is key for several regions of the world that have limited gauging and that are simultaneously facing significant threat due to anthropogenic climate change.

Referring back to Chapter 4, several elements of the catchment that were then assumed stationary, given that they varied only spatially and the focus was on developing a viable model at the catchment scale, are now to be considered non-stationary; if a model is to be developed that is able to accurately generalise across catchments, then spatial variables are of key concern. So, we extend the formulation of the model from being dependent only on climatic variables, $\mathbf{x}_c$, to catchment descriptor variables, $\mathbf{x}_d$, as well, s.t. the streamflow, y, is given by Equation 6.1:

$$y = f(\mathbf{x}_c, \mathbf{x}_d) \tag{6.1}$$

As assumptions had to be made about the climatic variable input, so too must similar assumptions be made about the catchment descriptors as inputs to the model, in terms of data availability and its veracity. For those catchments poorly described and less studied, again using the number of hydrological measuring stations reporting to extensive data platforms (The Global Runoff Data Centre, 2022) as a frame of reference, internal measurements will be considered unavailable; instead, the baseline will default to more expansive, earth observation datasets, which might also help to eliminate site-specific biases.

6.1 Topography, Land Use, & Geology

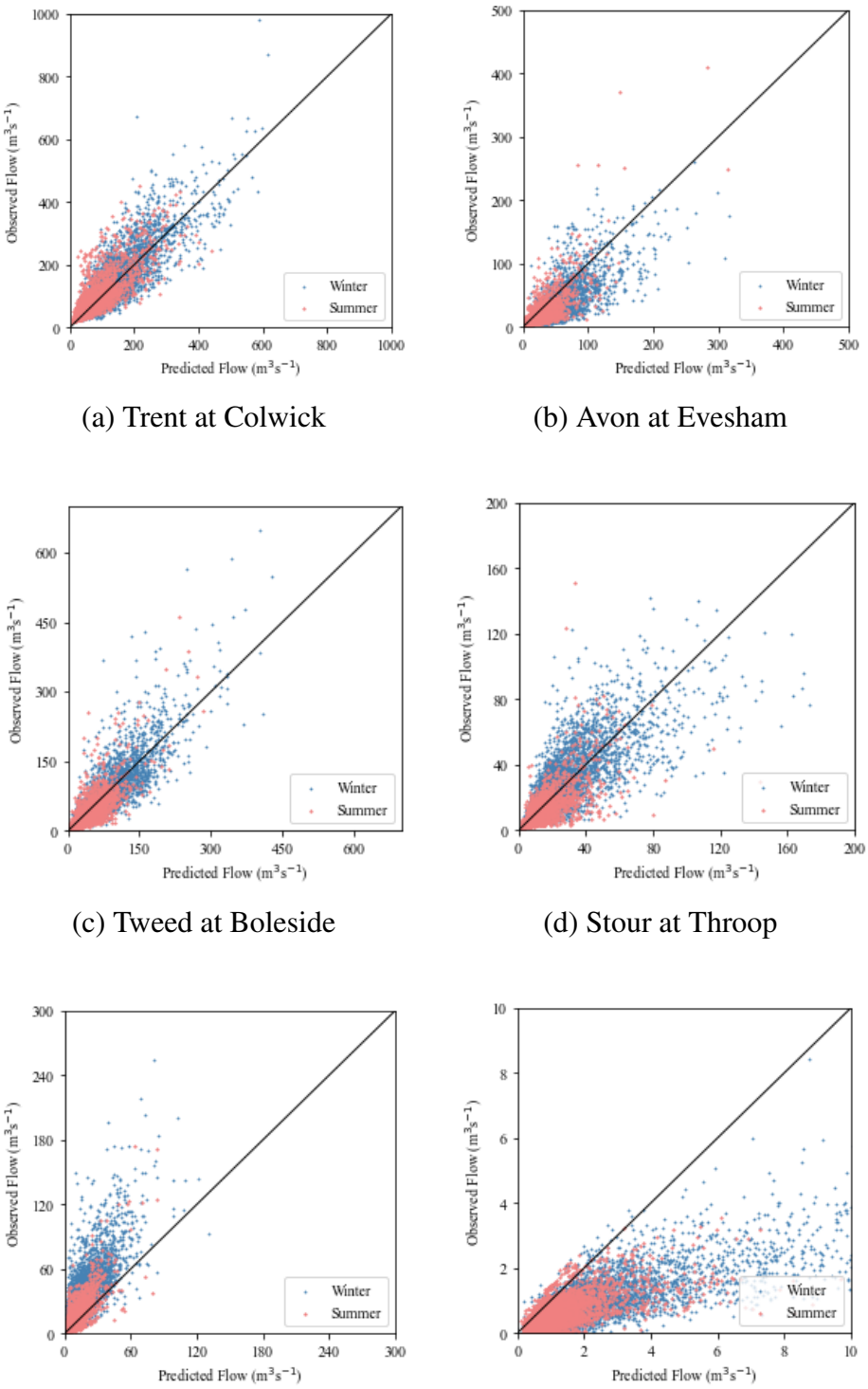
The guiding philosophy of this thesis has been to develop a modelling approach that is as lean as possible; thus, the Rational Model will serve as our starting point. This method is a paragon of simplicity and, yet, has proven useful for decades in successfully predicting peak discharges for urban drainage across the UK (Working Party on the Hydraulic Design of Storm Sewers and Hydraulics Research Limited, 1981). The formulation for the Rational Model is as follows:

$$y_P = c \cdot x_i \cdot A \tag{6.2}$$

Where y_P is the peak streamflow, c is a dimensionless coefficient, x_i is the average rainfall intensity for the period of interest, and A is the area of the catchment. If the dimensionless coefficient is equated to a weight within a neural network and the intensity of rainfall is subsumed into the climatic record from the single catchment models, then this initial model comprises just one new element for generalising across geographies: catchment area. The justification for this is via dimensional analysis; assuming an average rainfall over the catchment, then multiplication by its area gives the volume of water added to the catchment. If internal catchment effects, those processes that act to adjust the rate and location of water removal from said catchment, are assumed either constant or negligible between catchments, then a multi-catchment model based on area alone should deliver acceptable results.

A total of 124 catchments with complete records spanning 1979-2019, the date range for climatic variables extracted from ERA5, form the training and validation sets, whilst the test is comprised of 6 catchments, namely the River Trent at Colwick, the Avon at Evesham, the Tweed at Boleside, the Stour at Throop, the Exe at Thorverton, and the Darent at Hawley, taken as a random sample from the whole set with sampling rules to ensure a reasonable cross section of catchment areas (D. G. Morris and R. W. Flavin, 1990, 1994; UK Centre for Ecology & Hydrology, 2022).

The final step of the multi-catchment model formulation is the machine learning itself. Given the high performance, ease of deployment, and rapid training exhibited by the MLP8 in Chapter 5, the same model architecture is used here; apart from the expansion of the input space by one dimension to include the catchment area, the main difference being the size of the datasets, with over 1.5 million data points in the training and validation dataset. The results from training this MLP model are shown in Table 6.1 and Figures 6.1 and 6.2.



(a) Trent at Colwick

(b) Avon at Evesham

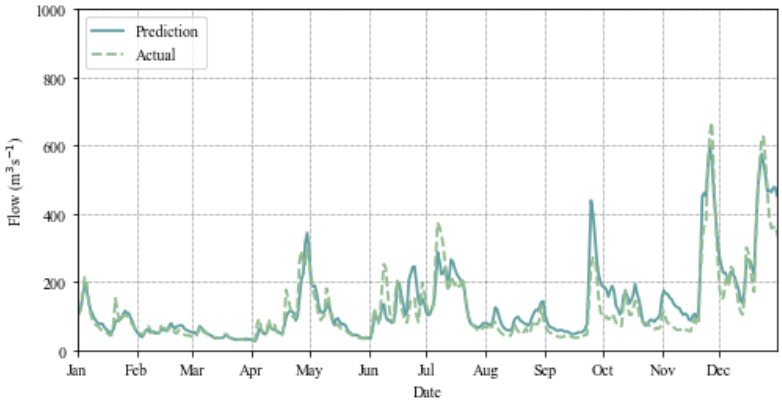
(c) Tweed at Boleside

(d) Stour at Throop

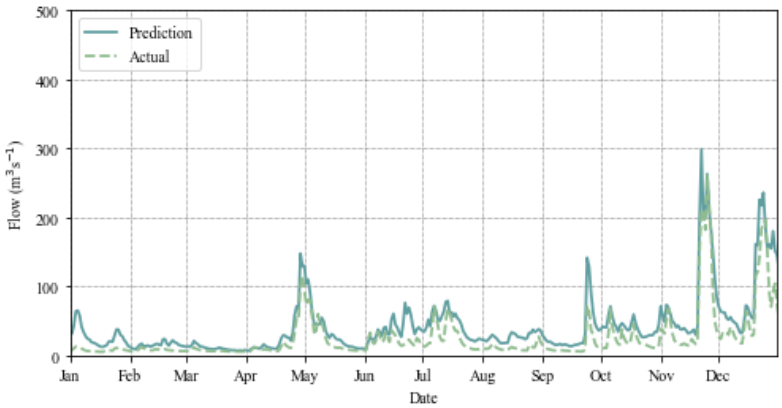
(e) Exe at Thorverton

(f) Darent at Hawley

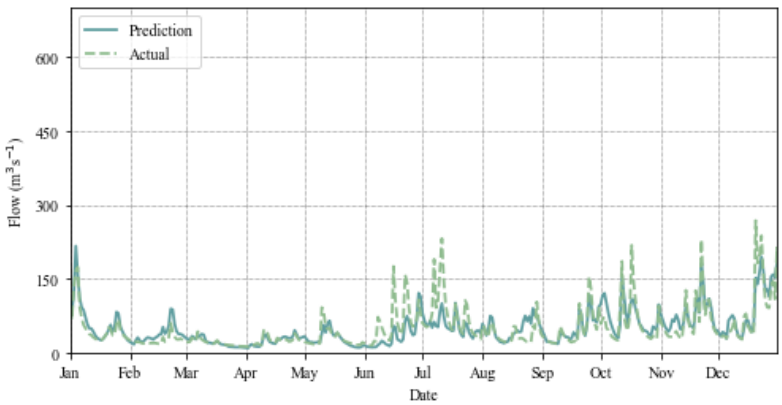
Fig. 6.1 Predictions against observations for the six test catchments using the area multi-catchment MLP



(a) Trent at Colwick

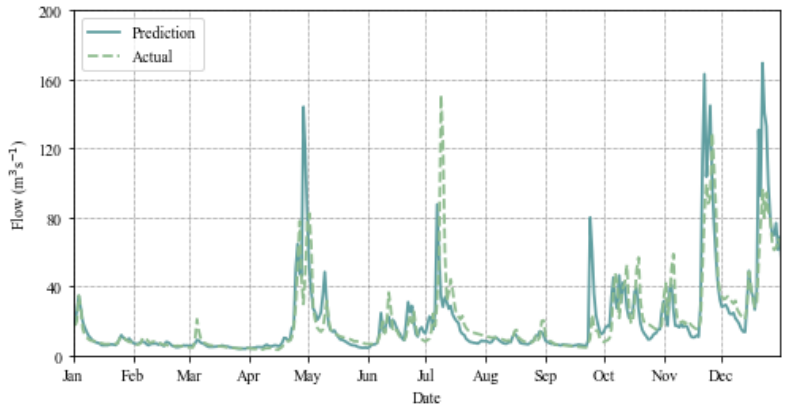


(b) Avon at Evesham

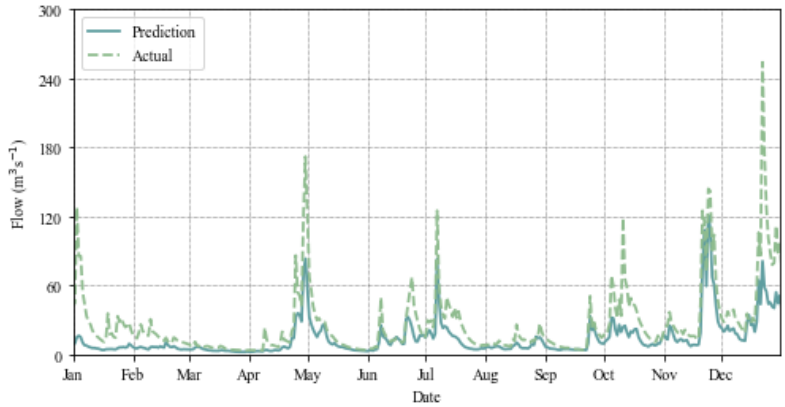


(c) Tweed at Boleside

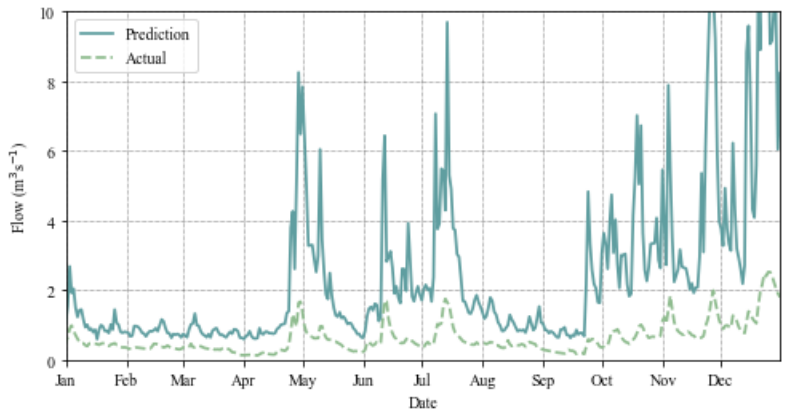
Fig. 6.2 Streamflow time series for the test set year, 2012, with predictions and observations using the area multi-catchment MLP



(d) Stour at Throop



(e) Exe at Thorverton



(f) Darent at Hawley

Fig. 6.2 Streamflow time series for the test set year, 2012, with predictions and observations using the area multi-catchment MLP

Table 6.1 Area multi-catchment MLP prediction performance by catchment

Catchment	Metric							
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})					
			$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Trent at Colwick	0.776	0.713	38.6	33.1	58.1	55.0	100	96.9
Avon at Evesham	0.357	0.489	6.40	10.0	9.47	16.9	17.7	32.0
Tweed at Boleside	0.773	0.678	13.3	16.3	24.5	28.2	48.5	53.4
Stour at Throop	0.684	0.636	3.93	4.72	7.77	7.68	17.0	15.7
Exe at Thorverton	0.451	-1.30	4.18	3.17	8.65	5.40	21.2	11.0
Darent at Hawley	-5.26	0.102	0.240	0.643	0.512	1.00	0.884	1.77
Average	-0.370	0.220						

Predictive performance across the test set is reasonable, to an extent, certainly being usable if the threshold for acceptable performance lies at an NSE of 0.5 for most catchments, with the major exception being the Darent at Hawley; performance, when weighted towards extremes through NSERA, also shows some promise. A similar approach adopted for ungauged streamflow prediction using artificial neural networks, albeit with different optimisation, within the same basin and, therefore, assuming catchment descriptors remained the same (Besaw et al., 2010) had performance below the level obtained here but was deemed to highlight the viability of the approach. That conclusion, in that the model could indeed be deemed satisfactory for industrial use, might be apt if the mean performance were not being so impacted by the model's poor predictive performance for the Darent. Closer inspection of the figures reveals that the streamflow for the Darent is being overpredicted by a significant amount, hence resulting in a large, negative NSE, and from which it is possible to infer that the internal processes, hitherto ignored, are attenuating, storing, or removing the majority of the precipitation over the catchment. Of course, the notion that internal processes have limited effect on the location and rate at which the rainfall added to a catchment is removed was overly simple and incorrect but it does provide a point from which to pivot.

At first refinement, the catchment descriptors that are now considered inputs for the model are those that affect the mass-balance equation from a more natural perspective acting to attenuate or amplify rainfall response. Not considered are the more direct forms of human influence, such as abstraction, effluent discharge, and so on, on streamflow. The data sources for these descriptors is given in Chapter 3 but we now provide further rationale behind their inclusion.

The topography of a catchment, specifically the two dimensional elevation data as distributed across a catchment domain, once compressed into a one dimensional statistical distribution can provide the average slope function; if combined with the radius, by approximating the area as a circle, a two dimensional surface can be produced that represents the topography, albeit through very few data points. Our expectation is that the machine learning algorithms at play will be able to represent this internally.

The second subset is that which describes land use, broken down into the proportion of the catchment with arable/horticultural, grassland, mountain/heath/bog, urban, or woodland covers. The different types of land cover will affect the rate at which water flows into a river, through surface and subsurface flows, in addition to the amount lost to evapotranspiration. Land use will encode a certain amount of anthropogenic influence on a catchment, given that the amount of urban cover will be correlated with the local population to a certain extent but does not reflect local population density. In other words, the indirect effect of urban cover, through man-made materials altering surface and subsurface flow, will be represented but human activity, such as abstraction, is likely to be different for catchments based on socio-demographic information not represented, for example a catchment feeding much of the Thames, with London's significant population density, as say compared to Devon (Park, 2021).

Finally, the third subset is the geology of the catchment, affecting the subsurface flows, such as percolation or groundwater recharge, as described according to categorisation by the British Geological Survey 1:625000 Bedrock Geology layer (British Geological Survey, 2007). A catchment is thus described by the proportional makeup of its underlying rock formations in terms of permeability with: high being that with highly productive fissured aquifers or aquifers with intergranular flow; moderate being that with locally important fissured aquifers or aquifers with intergranular flow; low being that with impermeable rock and negligible groundwater beyond soil depth; and mixed being that with concealed aquifers or those that are limited or only locally relevant.

Thus the set of catchment descriptors is now:

- Catchment area;
- Minimum elevation above sea level;
- 10th percentile elevation above sea level;
- 50th percentile elevation above sea level;
- 90th percentile elevation above sea level;
- Maximum elevation above sea level;
- Proportion of catchment area with arable/horticultural coverage;

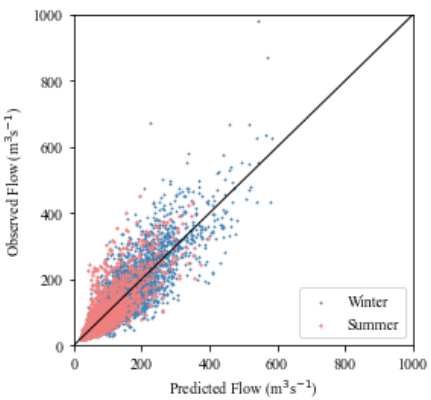
- Proportion of catchment area with grassland coverage;
- Proportion of catchment area with mountain/heath/bog coverage;
- Proportion of catchment area with urban coverage;
- Proportion of catchment area with woodland coverage;
- Proportion of bedrock that has high permeability;
- Proportion of bedrock that has medium permeability;
- Proportion of bedrock that has low permeability;
- Proportion of bedrock that has mixed permeability;

Many of the catchment descriptors used here, although provided directly through the NRFA, could be readily obtained though other means, such as using satellite imagery as provided by the Copernicus Global Land Service (Buchhorn et al., 2020) at resolutions up to 100m. Using the NRFA datasets expedites the upstream data ingress pipeline for this preliminary study covering catchments within the United Kingdom; however, the Copernicus Global Land Service provides extensive, global coverage, as the name might suggest, and would therefore be suitable for expansion of the model to encompass training and test data from across the world after its validation.

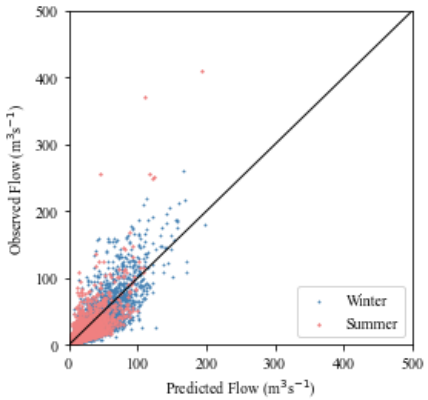
Table 6.2 Topography, land use, & geology multi-catchment MLP prediction performance by catchment

Catchment	Metric							
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})					
			$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Trent at Colwick	0.799	0.709	38.6	45.9	58.1	65.8	100	105
Avon at Evesham	0.632	0.468	6.40	9.19	9.47	14.6	17.7	25.4
Tweed at Boleside	0.773	0.624	13.3	16.6	24.5	27.7	48.5	50.0
Stour at Throop	0.729	0.527	3.93	3.28	7.77	5.60	17.0	12.6
Exe at Thorverton	0.642	-0.017	4.18	3.16	8.65	6.33	21.2	13.9
Darent at Hawley	-0.488	-0.568	0.240	0.758	0.512	0.995	0.884	1.31
Average	0.515	0.291						

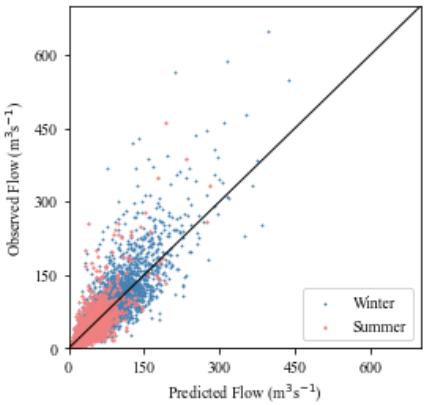
The expanded set of catchment descriptors results in superior model performance for all catchments, Table 6.2 and Figures 6.3 and 6.4 and whilst that does include the Darent at Hawley, both NSE and NSE-RA continue to be in negative territory for that catchment. The model performance is generally acceptable, by the prior standard of having an NSE$>$ 0.5,



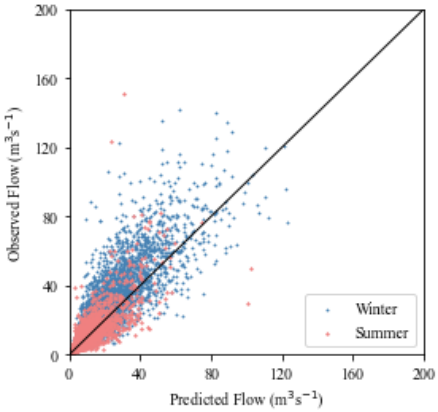
(a) Trent at Colwick



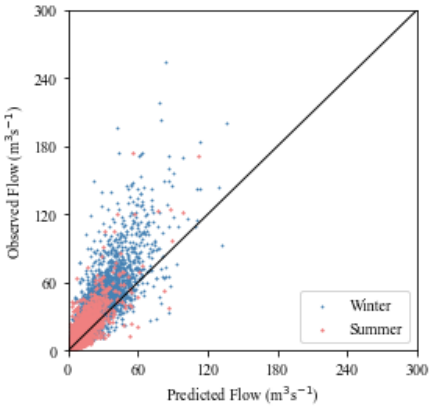
(b) Avon at Evesham



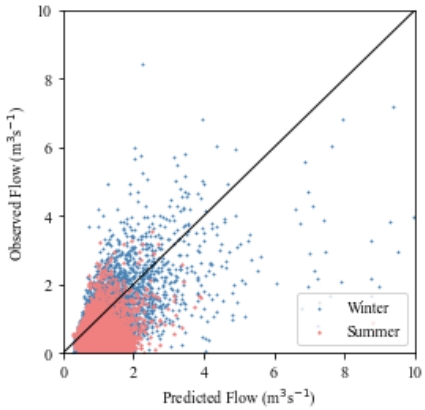
(c) Tweed at Boleside



(d) Stour at Throop

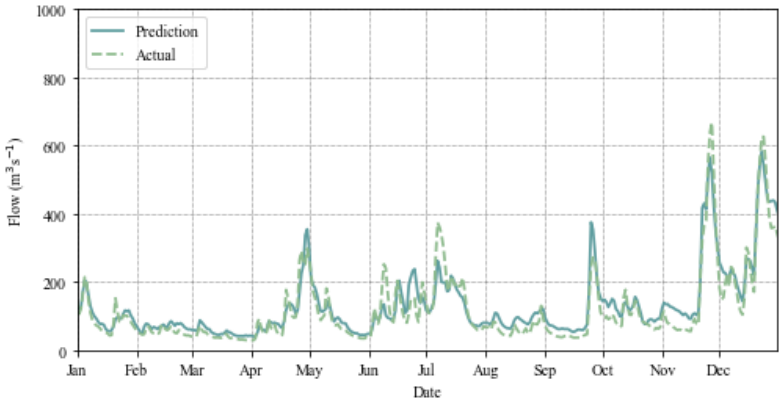


(e) Exe at Thorverton

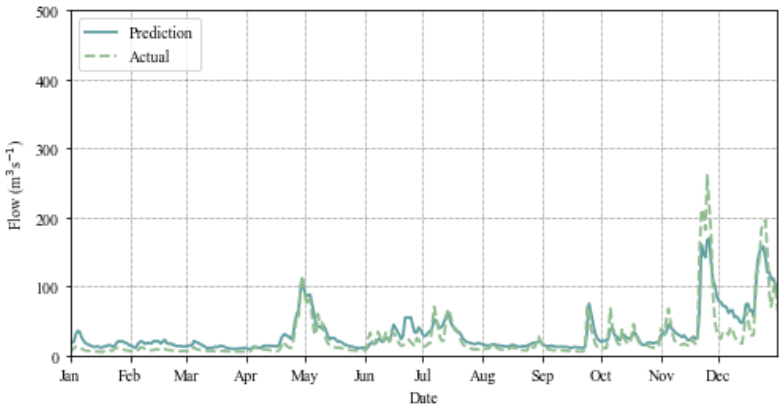


(f) Darent at Hawley

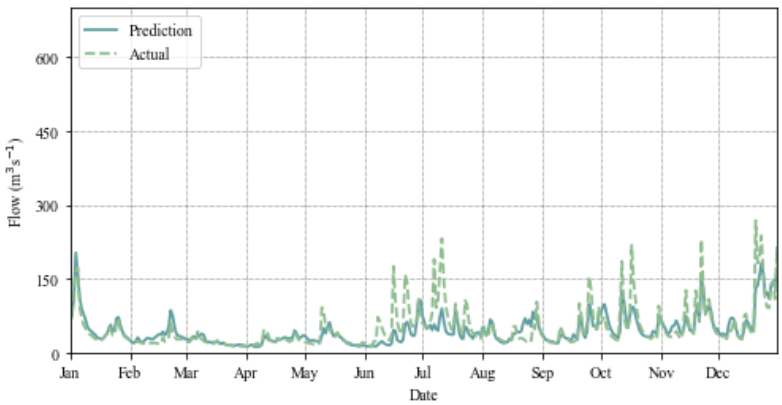
Fig. 6.3 Predictions against observations for the six test catchments using the topography, land use, & geology multi-catchment MLP



(a) Trent at Colwick

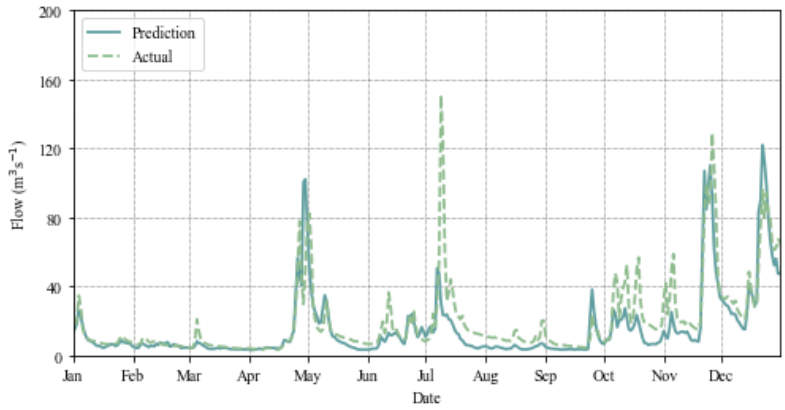


(b) Avon at Evesham

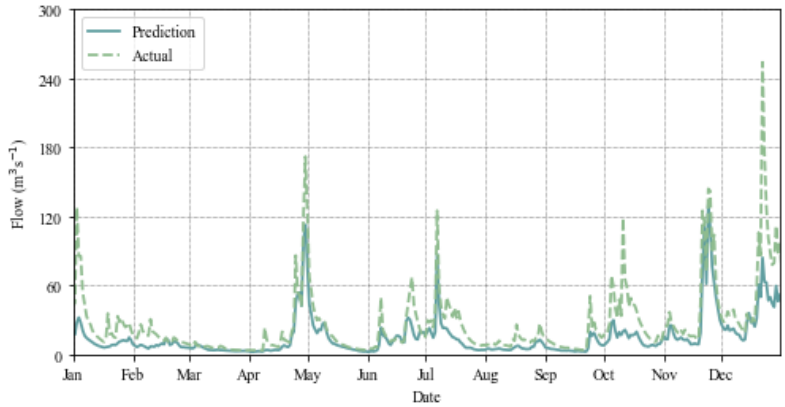


(c) Tweed at Boleside

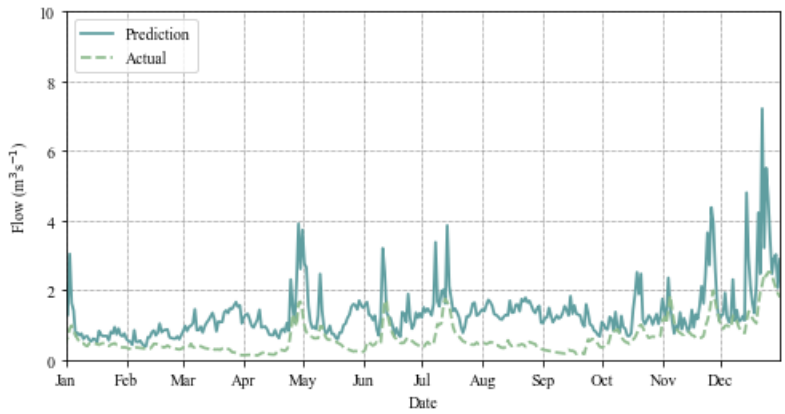
Fig. 6.4 Streamflow time series for the test set year, 2012, with predictions and observations using the topography, land use, & geology multi-catchment MLP



(d) Stour at Throop



(e) Exe at Thorverton



(f) Darent at Hawley

Fig. 6.4 Streamflow time series for the test set year, 2012, with predictions and observations using the topography, land use, & geology multi-catchment MLP

and the approach is at the moment similar to that used in a study focused on the United States (Kratzert et al., 2019), where the modelling approach used the LSTM form of RNNs to achieve similar performance; the mean NSE obtained in their study was almost identical to the one achieved here, at 0.715 when the Darent is dropped from the test set.

So if the model is to be further improved, then the catchment mechanics for the Darent at Hawley must be explored; whilst the MSE loss was noted in the study by Kratzert et al. to be suboptimal, given that it will result in a better fit for larger catchments with larger streamflows and subsequent prediction error, their refinement offered a small improvement to NSE of 0.05 and is, thus, not the issue here. Interestingly, the other catchment for which performance is problematic, albeit less so than the Darent, is The Exe at Thorverton; the station at Thorverton is downstream of the Pixton and the streamflow at which therefore captures a larger part of the Exe's total catchment area.

6.2 Approximating Human Behaviour

On further inspection, results for the test catchments, and even the fit for those catchments that make up the training set, reveal either general over- or under-prediction based on whether or not the catchment, according to labelling in the NRFA, is either close to being a natural system or not. More concretely, the flow for those catchments with human intervention is over-predicted and the flow for those catchments without human intervention is under-predicted. It would appear that for all of the catchment descriptors used in the prior models of this section, that human behaviour is not being properly accounted for, even with land use characteristics being utilised.

Human behaviours that influence a catchment's behaviour can be broadly thought of as those that either remove or add water into a catchment's discharge; the problem is that the system is not necessarily closed under these operations. Consider the Darent at Hawley: this river is of note not only because it comprises part of the test set but because it is an exemplar of this problem. Sufficiently large volumes of water, with licences granted for up to 107 million litres per day, were abstracted from the Darent's catchment by the larger area water company, Thames Water, for domestic consumption, resulting in the river often running dry in the spring and summer (Willis, 1995); given that this corresponds to approximately $1.28m^3s^{-1}$ of abstraction and the mean streamflow over the period of interest is $0.5m^3s^{-1}$, the abstraction volume could feasibly represent up to two thirds of all water entering the catchment. Whilst a scheme to improve flow in the Darent was implemented to reroute water during dry conditions to maintain river ecology (Jones et al., 2012) following realisation of

the problematic abstraction volumes, the fact remains that this catchment has experienced significant disruption due to human influence.

The other side of this coin is what humans add to a river: industrial, agricultural, and residential effluence. Even within the tight regulatory environment of the EU, 60% of rivers in Europe have failed ecological health standards (Kristensen et al., 2018) with the UK being a particularly poor performer, where, out of 4679 surface water bodies in the UK, all were deemed to have failed regulatory standards for chemical pollution and only 16% have good ecological health (The Environment Agency, 2022). In 2019, English water companies discharged untreated human waste into streams and rivers for more than 1.5 million hours over 200,000 occasions with no monitors at outflows measuring the total discharge (Laville and McIntyre, 2020). Ignoring the harm caused by discharge of pollution and waste, the fact that this is unmeasured is in and of itself problematic. For example that resulted in a court case and a guilty plea, the evidence still didn't have concrete data; Southern Water Services Limited was found to be discharging waste at 17 sites into controlled coastal waters over 61,704 hours but a formal estimate of $0.850 m^3 s^{-1}$ was provided at only one site (Johnson, 2021). Extrapolating from a single data point, even though that was done in the court case, could be unwise and this is but one facet of human influence, one stroke within a much broader picture.

Ultimately, the assumption that these systems are closed under human interaction over a given time period should be discarded, due to water being stored or even moved between catchments; in some cases, such as the Darent, this volume can be highly disruptive. So, one ought to conclude that human interactions are having material effect and that they present a challenge in terms of measurement, either being missing, estimates, or highly sensitive; furthermore, if, for certain economies, streamflow and rainfall aren't extensively gauged, then it would be fair to assume that human interactions are even less so, given the comparative lack of data for an area such as the UK with its extensive river monitoring.

Thus, another proxy variable set is required, one that is simple enough to be transferable to areas with no viable data collection strategies, that is capable of expressing human interaction, and enables machine learning models to generate relatively robust predictions for streamflow in the face of said human interaction. The strategy offered in the face of this challenge is to utilise qualitative assessments, such as those provided through analysis conducted by CEH, on whether or not a catchment is affected by abstraction, effluent discharge, unnatural flow regulation, and storage (UK Centre for Ecology & Hydrology, 2022) to create four "quasi-binary" variables; these variables take values from the set $\mathscr{B} = \{0, 1, 2\}$, with 0 representing no indication of the influencing behaviour, 1 representing some indication of the influencing

behaviour, and 2 representing substantial indication of the influencing behaviour. So for our prime offender, the Darent at Hawley, one would set $x_{abstraction} = 2$.

The experimental setup is retained, so the MLP architecture and design remain broadly the same, as does the train/test split, and the overall training and tuning regime. The only change introduced is the expansion of the feature set to include the input variables that represent the above quantitative interpretation of the qualitative assessment provided by CEH. The impact on performance through including these variables is shown in Table 6.3.

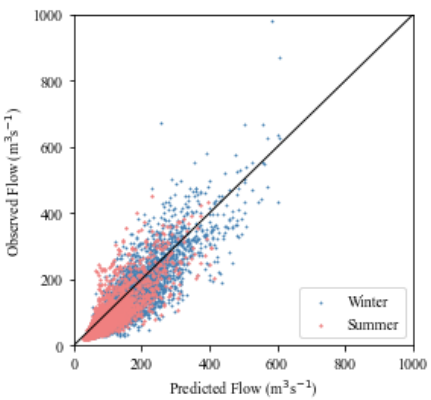
Table 6.3 Anthropogenic multi-catchment MLP prediction performance by catchment

Catchment	Metric							
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})					
			$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Trent at Colwick	0.720	0.668	38.6	60.0	58.1	82.0	100	126
Avon at Evesham	0.651	0.180	6.40	6.7	9.47	10.9	17.7	19.0
Tweed at Boleside	0.786	0.695	13.3	18.9	24.5	29.6	48.5	52.7
Stour at Throop	0.672	0.351	3.93	2.45	7.77	4.73	17.0	10.7
Exe at Thorverton	0.775	0.572	4.18	4.26	8.65	8.65	21.2	18.3
Darent at Hawley	0.416	0.169	0.240	0.173	0.269	1.00	0.461	1.77
Average	0.670	0.439						

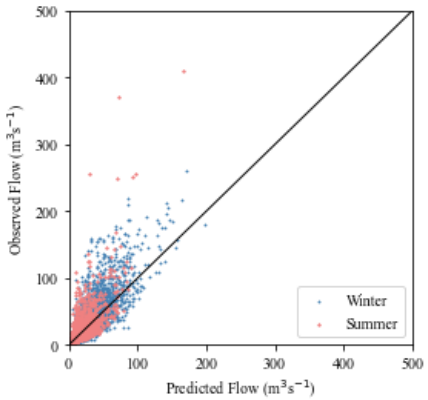
The inclusion of the quasi-binary human influencing behaviour variables delivers a substantial improvement in the mean performance both generally and for extreme values, with the improvement in the predictive capability for the Darent at Hawley, as is shown in Figures 6.5 and 6.6, driving the average for both NSE and NSE-RA higher. Whilst NSE is improved for the less influenced catchments, there is a slight decrease in NSE-RA for those same catchments.

The sensitivity analysis approach, also from Chapter 4, is adopted here, perturbing the network by the maxima and minima for each of the input variables, with the resulting sensitivities for catchment descriptors and antecedent proxies averaged across the six test catchments shown in Table 6.4 and climatic variables represented graphically in Figure 4.7 for comparison.

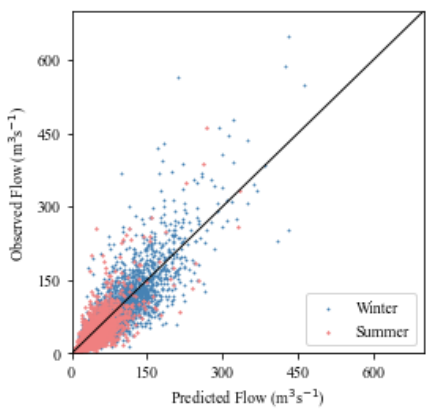
Given the evolution of the model to this point, with the Rational approach providing the starting point and adequate performance for most catchments, and as borne out through the sensitivity analysis, the relative importance of catchment area is evident, being the single



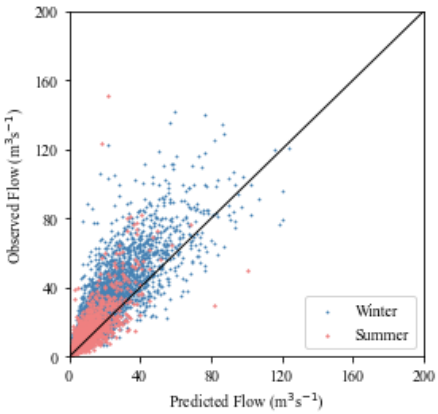
(a) Trent at Colwick



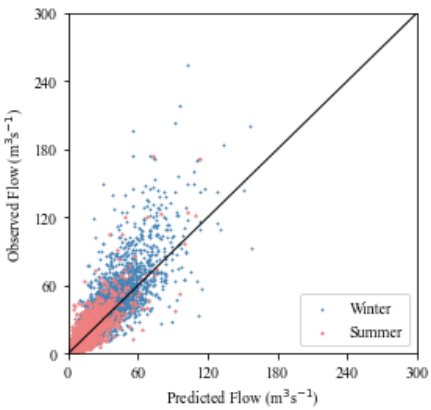
(b) Avon at Evesham



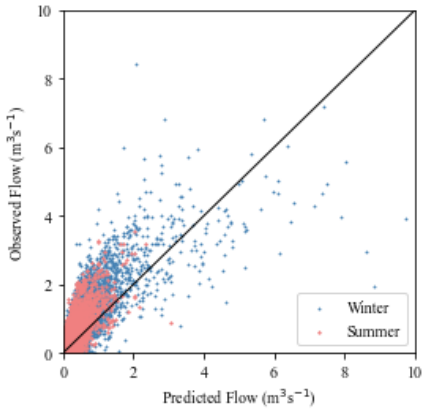
(c) Tweed at Boleside



(d) Stour at Throop

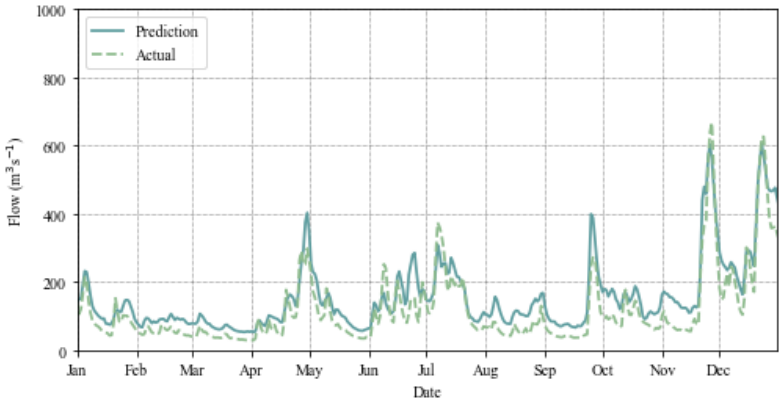


(e) Exe at Thorverton

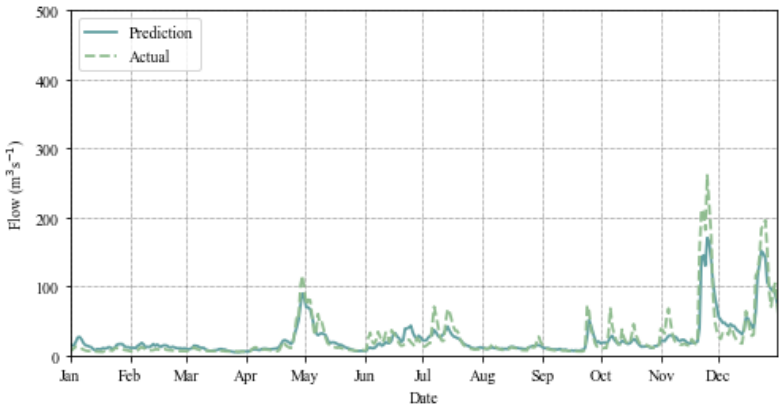


(f) Darent at Hawley

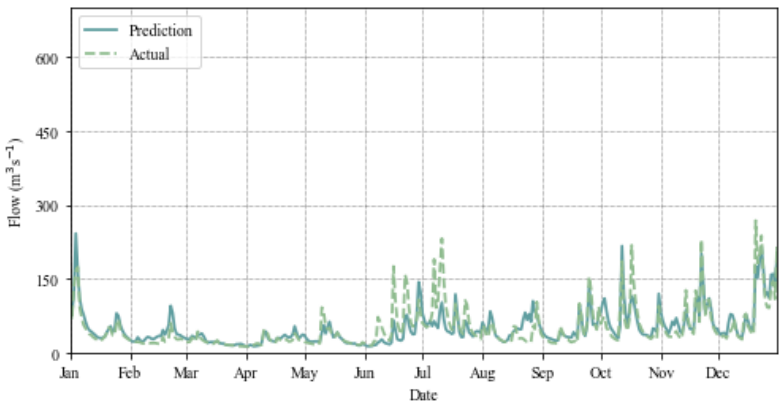
Fig. 6.5 Predictions against observations for the six test catchments using the anthropogenic multi-catchment MLP



(a) Trent at Colwick

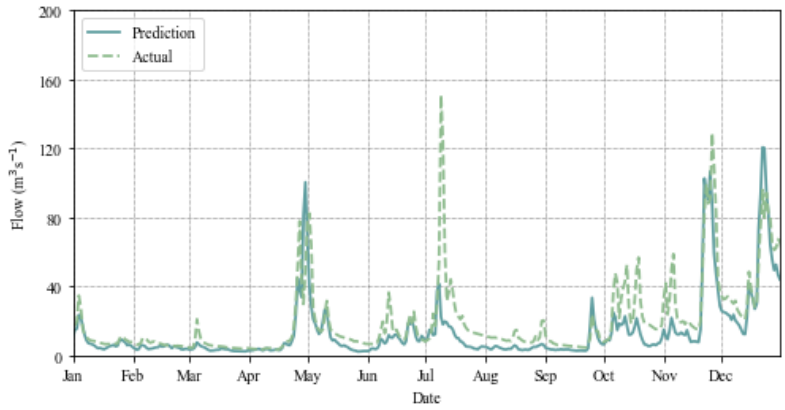


(b) Avon at Evesham

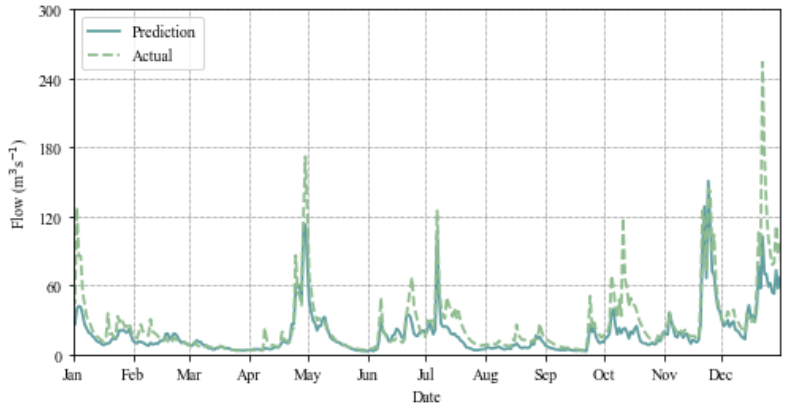


(c) Tweed at Boleside

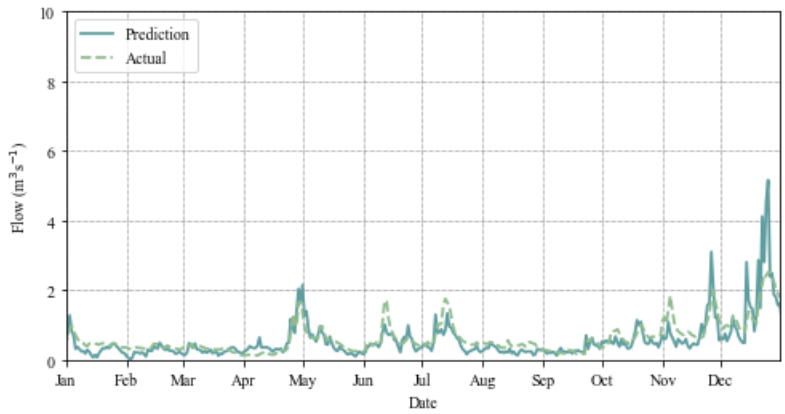
Fig. 6.6 Streamflow time series for the test set year, 2012, with predictions and observations using the anthropogenic multi-catchment MLP



(d) Stour at Throop



(e) Exe at Thorverton

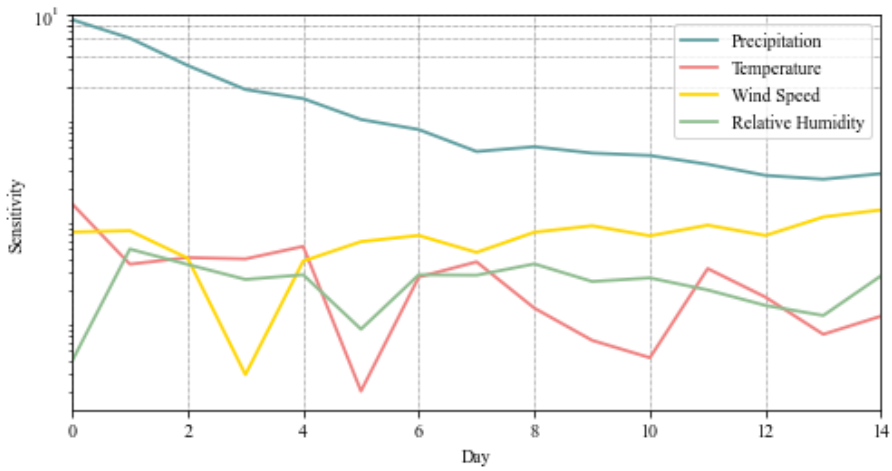


(f) Darent at Hawley

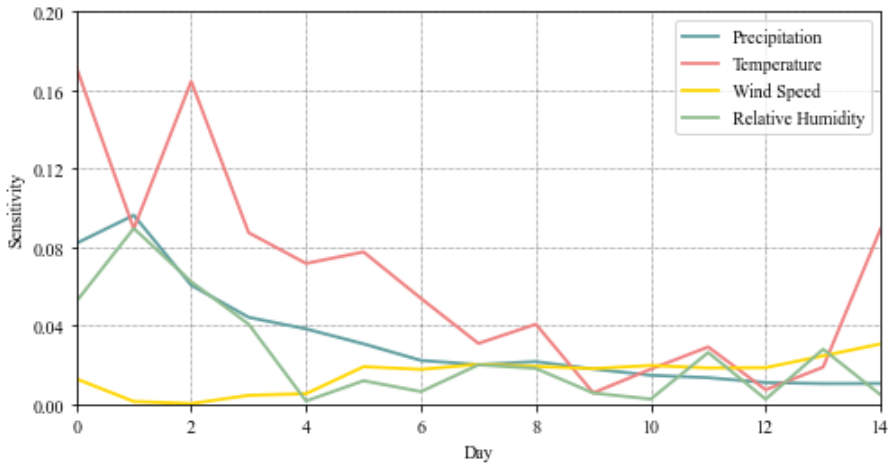
Fig. 6.6 Streamflow time series for the test set year, 2012, with predictions and observations using the anthropogenic multi-catchment MLP

Table 6.4 MLP Sensitivity to the maximum and minimum perturbations for each catchment descriptor variable and antecedent proxies

Variable	Sensitivity	
	Perturbation$_{minimum}$	Perturbation$_{maximum}$
Catchment area	0.8961	8.6374
Minimum elevation	0.0423	0.1417
10th percentile elevation	0.1371	0.2918
50th percentile elevation	0.3350	0.8765
90th percentile elevation	0.1780	0.4283
Maximum elevation	0.0782	0.0777
Arable/horticultural coverage	0.0498	0.0373
Grassland coverage	0.2722	0.2184
Mountain/heath/bog coverage	0.3226	0.6747
Urban coverage	0.1925	0.6907
Woodland coverage	0.0619	0.2313
High permeability	0.0373	0.3346
Medium permeability	0.0423	0.4597
Low permeability	0.0242	0.1196
Mixed permeability	0.0301	0.5421
Abstraction impact	0.0684	0.3329
Effluence impact	0.0076	0.0538
Regulation impact	0.0210	0.2829
Storage impact	0.0586	0.0104
Natural vs Influenced Impact	0.0448	0.1698
FARL	0.8959	0.0716
1 Month average precipitation	0.2664	1.7375
3 Month average precipitation	0.3036	1.2539
6 Month average precipitation	0.3699	0.4223
1 Month average temperature	0.4753	0.3059
3 Month average temperature	1.6105	0.6436
6 Month average temperature	0.6119	1.3195



(a) Network response to climatic variable maximum



(b) Network response to climatic variable minimum

Fig. 6.7 Average network sensitivity to each climatic variable at time, t, for high input signal, top, and low input signal, bottom

most important catchment variable; the total amount of precipitation that a catchment is subjected to is a function of its area, so this result is expected.

Underscored again, as with the analysis in Chapter 4, is the relative importance of high rainfall and high temperatures. The proxy antecedent condition variables also have high bearing on the model's output, with the long term averages for rainfall becoming less so over time. If this trend were extrapolated, then perhaps increasing the length of time over which the average is calculated would see the impact decrease, as the relative weight of near term precipitation decreases. Conversely, the decreasing relevance is not seen with the long term averages of temperature, which may be down to length of time over which seasonal

influences act. Increasing the period to eliminate season variation would, in effect, reduce the impact to a climate change signal with the variable, assuming climate change non-stationary, then representing the increasing trend in national temperature.

The impact of some of the anthropogenic influence proxy variables is of a similar order of magnitude to the the land use and geological variables, further underscoring their utility; although the effect on performance was most notable for the Darent at Hawley, these variables are powerful, yet simple, and will be used in the subsequent sections of this chapter.

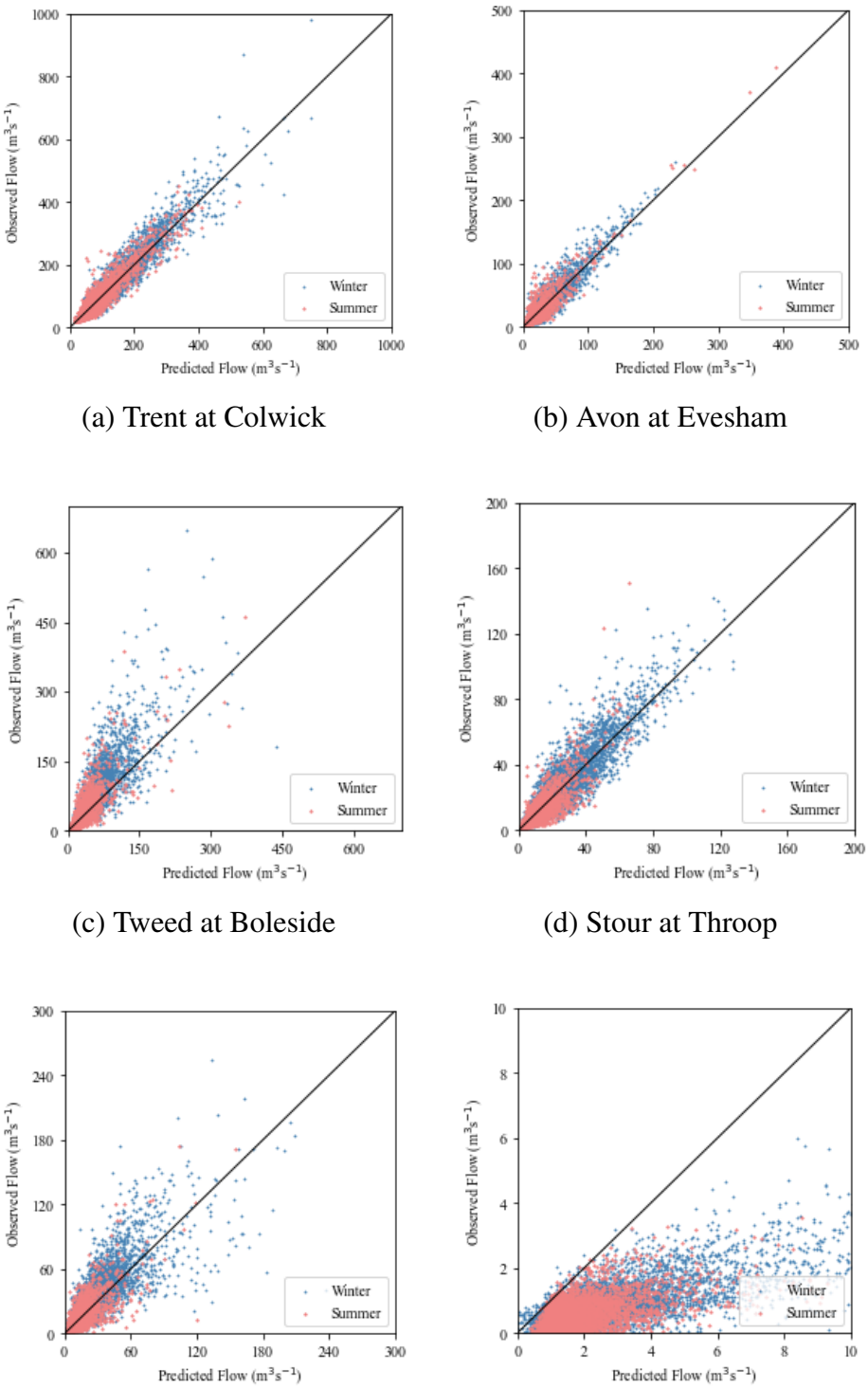
A potential weakness of this form of sensitivity analysis, however, is that it does not identify which features are encoding similar information and, therefore, it is possible that some features are potentially duplicating information encoded by others and that an approach to do so would lead to a leaner model without sacrificing accuracy (Chen et al., 2020). For example, it is likely that, given they are proportional variables, the geological catchment descriptors are an example of this, whereby the proportions of bedrock type sum to 1 and by providing only 3 of 4 the value of the 4th is necessarily encoded through their combination. Even so, all variables included thus far appear to form a suitable model approach for a generalising hydrological model.

6.3 An Ensemble Approach

Rather than investigating the full complement of models from Chapter 5, the most consistent, best performing models, deep MLPs, NPs, and GPs will be utilised to determine which method is the most suitable for this wider application. The deep MLP was already used in the preceding sections of this chapter, with the best obtained performance in Table 6.3. The implementation of the NP here is of similar structure to that from Chapter 5, with the decoder and encoder networks matching the structure of the anthropogenic MLP. The implementation of a GP, however, becomes somewhat problematic due to specification of the kernel and the size of the dataset; although the SVGP model is suitable for a dataset of this size, but the suitability of the kernel over the catchment descriptors is less obvious. Given the ubiquity of the Squared-exponential kernel, this will be used in conjunction with the kernel from Chapter 5 over the climatic variables.

Generating predictions using the NP and GP are shown in Tables 6.5 and 6.6, respectively.

Even with the quasi-binary variables that improved the MLP for the Darent, prediction for that river remains problematic for both the NP and the GP. With regards to the latter, the model is failing to account for the abstraction taking place within the Darent catchment, evidenced by the significant over prediction. Either the kernel is improperly specified or the SVGP method is unable to generalise well with the dimensionality of the problem. Methods



(a) Trent at Colwick

(b) Avon at Evesham

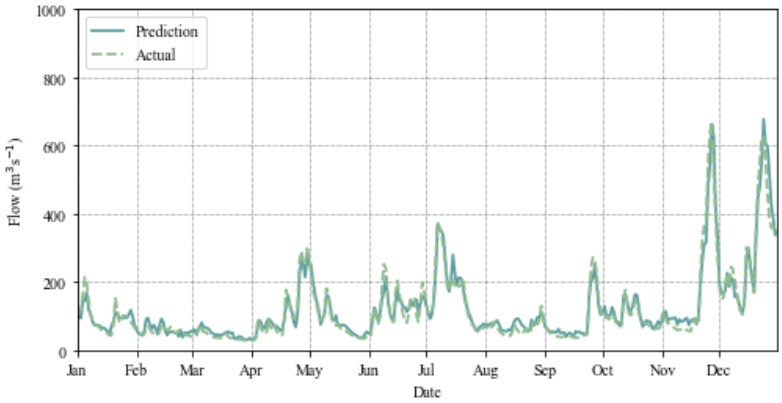
(c) Tweed at Boleside

(d) Stour at Throop

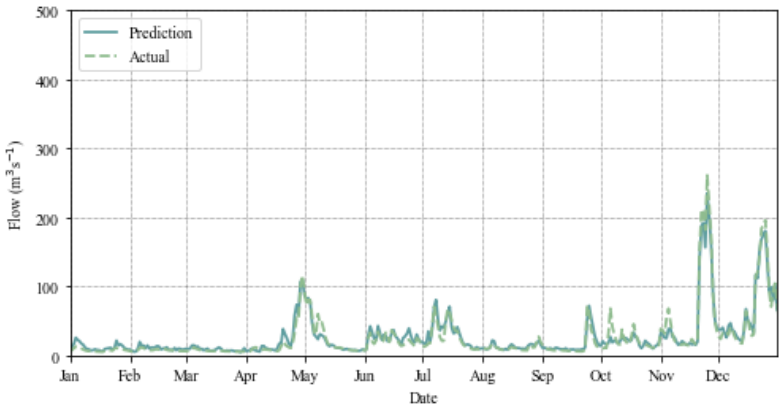
(e) Exe at Thorverton

(f) Darent at Hawley

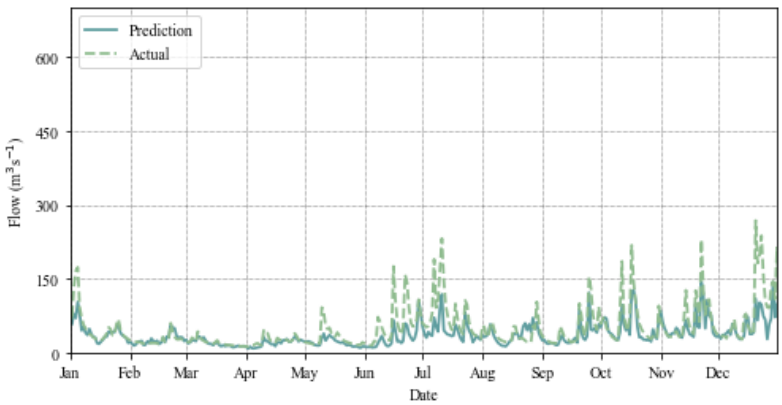
Fig. 6.8 Predictions against observations for the six test catchments using the anthropogenic multi-catchment NP



(a) Trent at Colwick

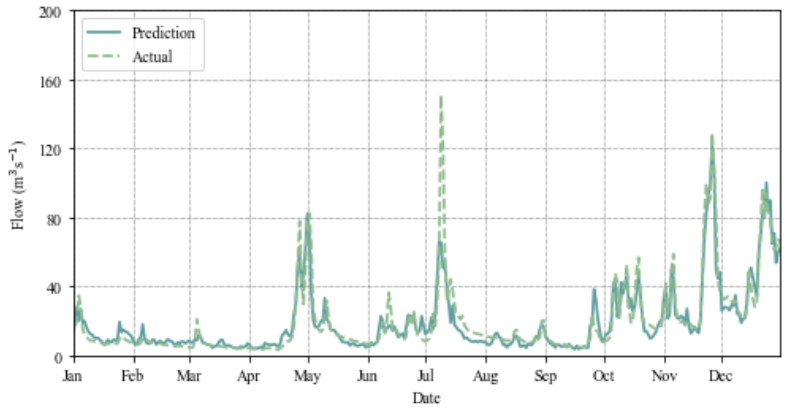


(b) Avon at Evesham

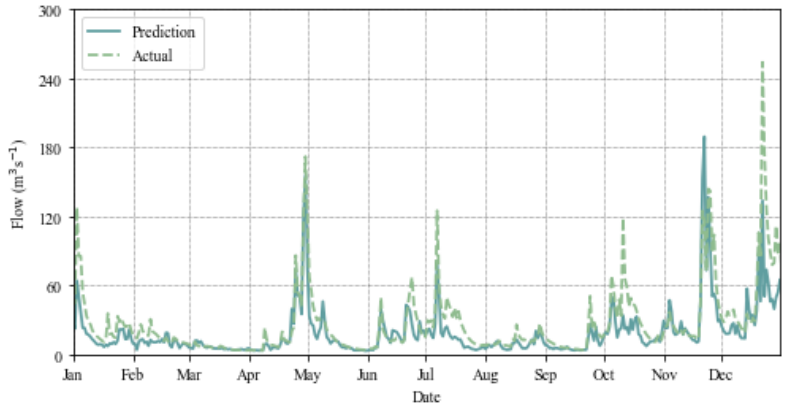


(c) Tweed at Boleside

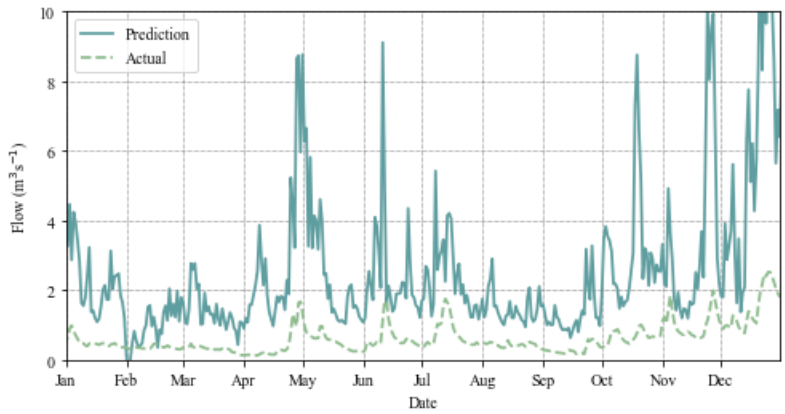
Fig. 6.9 Streamflow time series for the test set year, 2012, with predictions and observations using the anthropogenic multi-catchment NP



(d) Stour at Throop

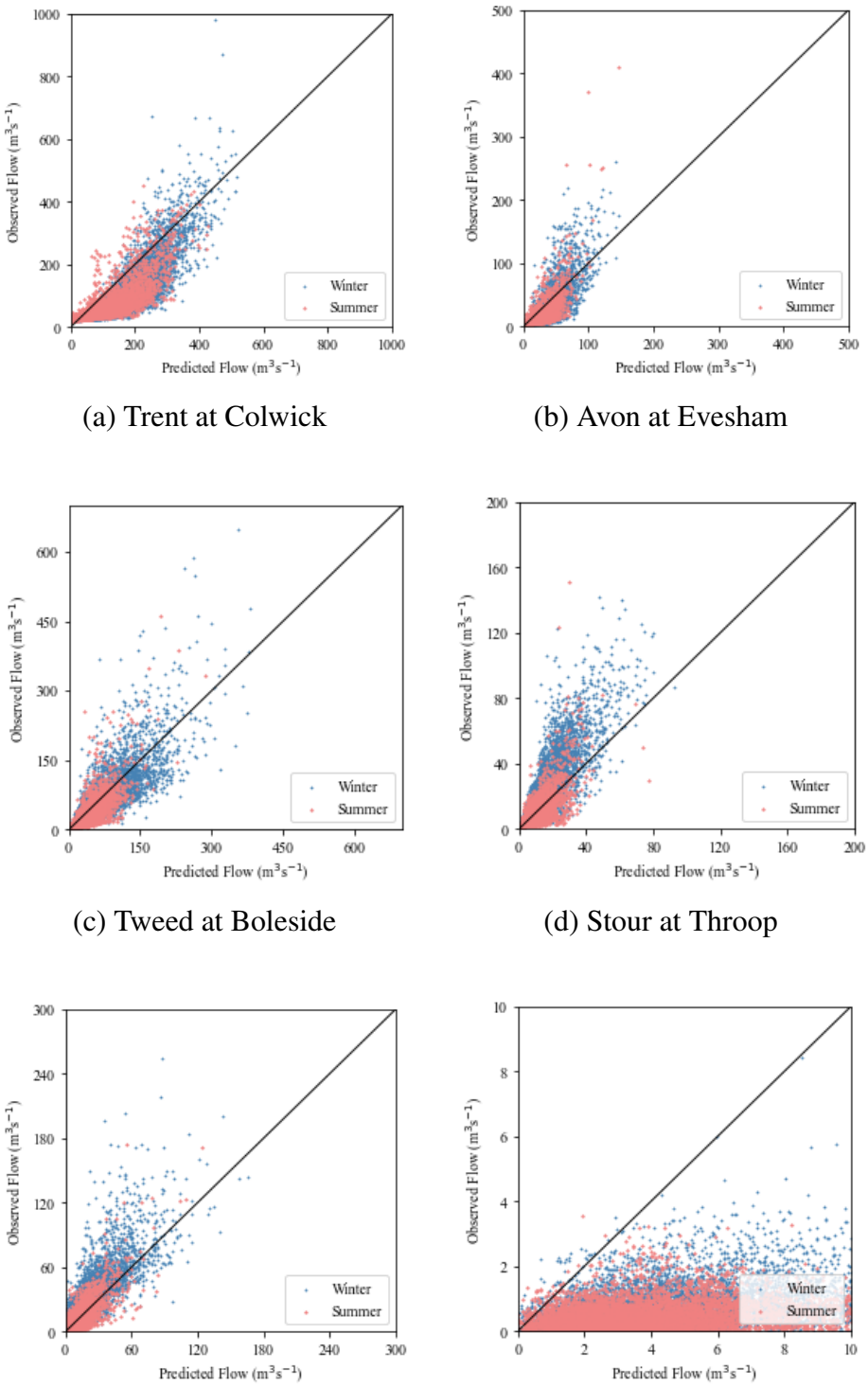


(e) Exe at Thorverton



(f) Darent at Hawley

Fig. 6.9 Streamflow time series for the test set year, 2012, with predictions and observations using the anthropogenic multi-catchment NP



(a) Trent at Colwick

(b) Avon at Evesham

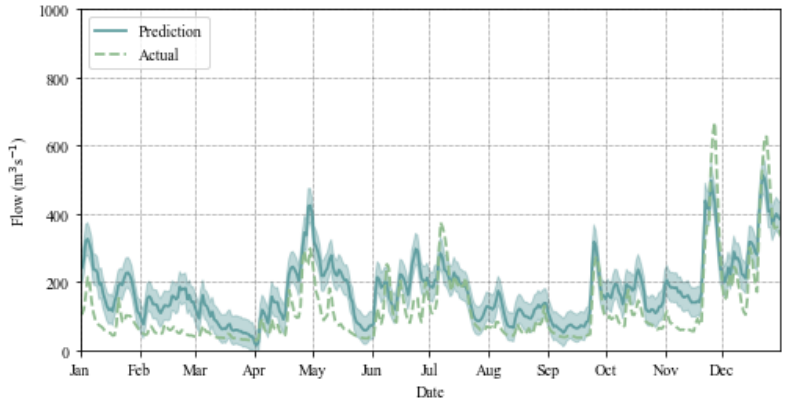
(c) Tweed at Boleside

(d) Stour at Throop

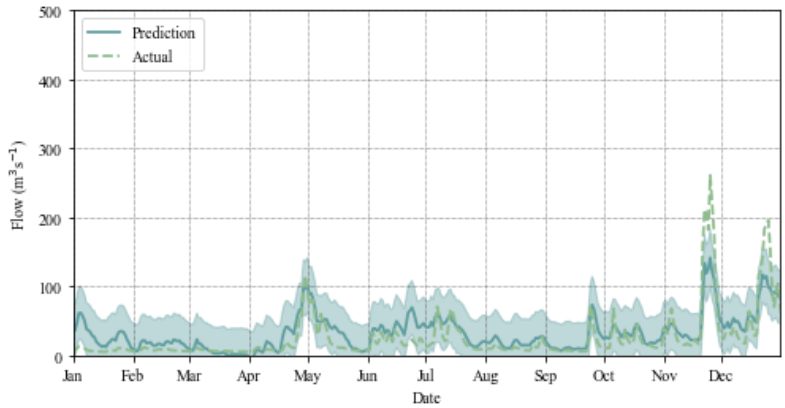
(e) Exe at Thorverton

(f) Darent at Hawley

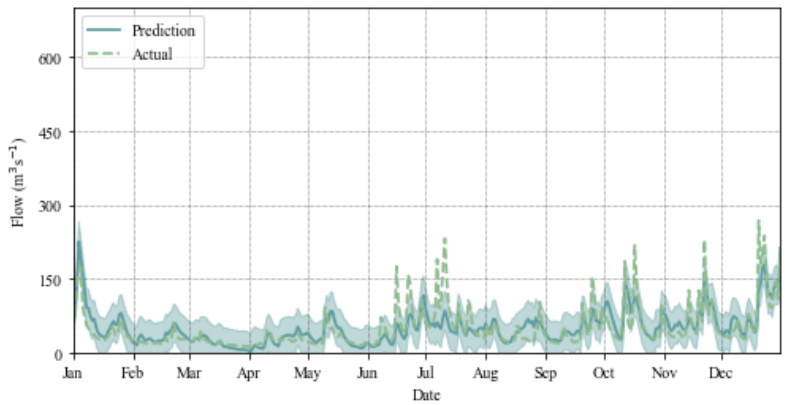
Fig. 6.10 Predictions against observations for the six test catchments using the anthropogenic multi-catchment SSVGP



(a) Trent at Colwick

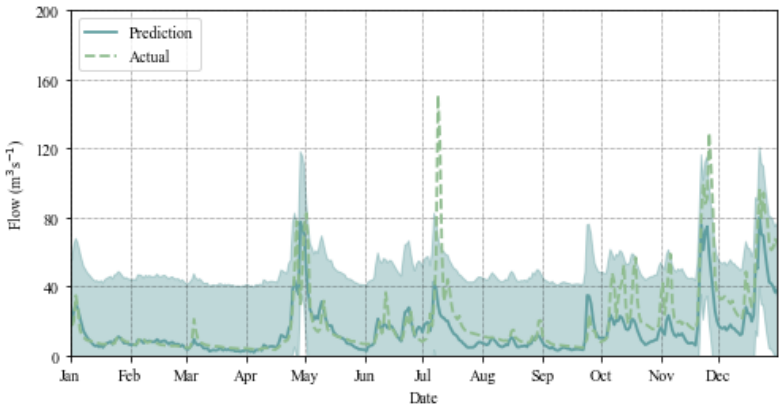


(b) Avon at Evesham

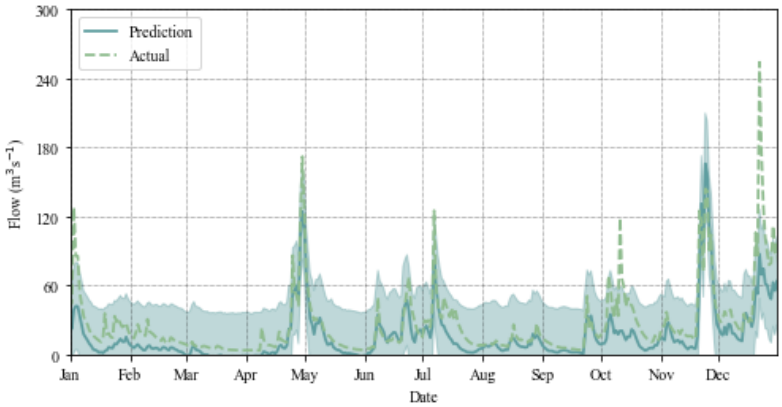


(c) Tweed at Boleside

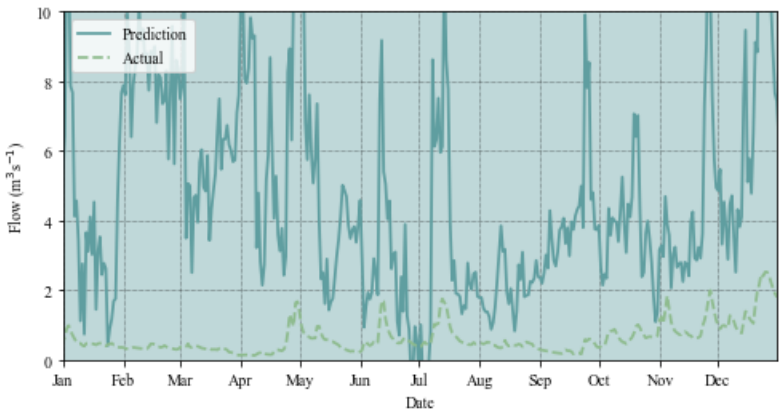
Fig. 6.11 Streamflow time series for the test set year, 2012, with predictions and observations using the anthropogenic multi-catchment SSVGP



(d) Stour at Throop



(e) Exe at Thorverton



(f) Darent at Hawley

Fig. 6.11 Streamflow time series for the test set year, 2012, with predictions and observations using the anthropogenic multi-catchment SSVGP

Table 6.5 Multi-catchment NP prediction performance by catchment

Catchment	Metric							
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})					
			$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Trent at Colwick	0.924	0.904	38.6	43.2	58.1	62.4	100	100
Avon at Evesham	0.892	0.869	6.40	6.97	9.47	10.3	17.7	17.9
Tweed at Boleside	0.669	0.485	13.3	14.0	24.5	22.2	48.5	37.7
Stour at Throop	0.869	0.821	3.93	4.42	7.77	8.41	17.0	16.9
Exe at Thorverton	0.722	0.589	4.18	4.36	8.65	8.03	21.2	17.0
Darent at Hawley	-10.3	-0.272	0.240	0.173	0.269	1.50	0.461	2.46
Average	-1.04	0.466						

Table 6.6 Multi-catchment SSVGP prediction performance by catchment

Catchment	Metric							
	NSE	NSE-RA	Flow Quantile (m^3s^{-1})					
			$Q_{25:O}$	$Q_{25:P}$	$Q_{50:O}$	$Q_{50:P}$	$Q_{75:O}$	$Q_{75:P}$
Trent at Colwick	0.703	0.382	38.6	34.4	58.1	67.7	100	111
Avon at Evesham	0.512	0.413	6.40	11.5	9.47	18.8	17.7	34.3
Tweed at Boleside	0.638	0.592	13.3	18.3	24.5	33.8	48.5	61.9
Stour at Throop	0.664	0.612	3.93	4.64	7.77	9.06	17.0	18.5
Exe at Thorverton	0.515	-0.802	4.18	2.50	8.65	6.38	21.2	13.5
Darent at Hawley	-47.5	-0.129	0.240	-1.29	0.269	1.13	0.461	4.16
Average	-1.04	0.499						

exist for automatic kernel design and for improving computational speed for ordinary GPs (Gardner et al., 2018; Steinruecken et al., 2019) that may make this model type easier to train and fit but that doesn't necessarily equate to suitability, particularly with regards to automatic kernel selection as this almost renders the GP without prior and akin to a more computationally expensive black-box model. Therefore, continuing with this model is not recommended, in spite of the positives that GPs had at the single catchment level, without further improving the kernel design for catchment descriptors; effort in this area, however,

may not necessarily be well spent, given the relative ease of use for the NP and its comparative strong performance.

For the NP, the model generalises well to all test catchments with the exception of the Darent, with performance similar across training catchments, so the model isn't necessarily overfitting the training data but, given the Darent is comparatively highly unusual, the ability of the NP to generalise to such extremes may be limited. And so a sensible strategy may be to combine the performance of the NP with a model that can generalise for these more unusual cases.

Thus, one can adopt an ensemble model approach, whereby the results from both the MLP and NP are used and weighted according to the level of human influence within a catchment. If the weighting is such that most catchments and rivers fall under the predictive purview of the NP, whilst those that have significant alterations to the flow regime due to human influences are more predicted by the MLP, then the average NSE across all test catchments is further increased, up to 0.758. If we consider that physically based distributed models in one study achieved an NSE of up to 0.400 on catchments which they were specifically calibrated to and the power of this methodology becomes apparent. Of course, it could be the case that, even with the large size of the training set, the number of catchments actually represented is insufficient for representing the set of all possible permutations of catchment variables; it follows that increasing the number of catchments within the database could benefit all of the modelling approaches to the point that the NP is able to perform across all catchments to the desired level and an ensemble approach is not required.

6.4 Global Expansion

There are certain scenarios that are likely not encountered by using a dataset comprised only of catchments within the United Kingdom, such as those experiencing significant snowfall or with glaciers, but machine learning models should be anticipated to be able to account for these variables. We also note that the list of easily obtained catchment descriptors was not exhaustive and perhaps factors such as the length of a river, its branching, and the shape of the catchment could be added into this framework with minimal computational or time cost.

Ergo, it should be anticipated that this method will have global applicability; the only requirement would be to obtain additional data from regions that encompass the types of flow regimes not found within the United Kingdom but that still come from areas of significant study, for example catchments that are bounded by mountainous regions in Europe or the United States of America.

Another modification that should be made, particularly with regards to the expansion of the framework to include developing countries, is for land and water use variables to change with respect to time. The assumption made for the United Kingdom, as a developed country, is that the amount of local development within a catchment over recent history has been limited and, therefore, the human influenced variables were deemed temporally stationary. Again referring to the Darent and as was noted before, the nature of abstraction within the catchment has varied with the Darent Augmentation Scheme being implemented within the timeframe of study but the effect on the streamflow as measured at Hawley was negligible. As shown in Figure 6.12, there is limited observable change over the time period for minimum, mean, and maximum flow by year with the increases, that occur in 2001 and 2014, coinciding with the catchment experiencing more rainfall on average in those years. However, the minimally observable impact of the Darent Augmentation Scheme, with regards to the data presented here, is not an effect that should be assumed to occur with every human intervention.

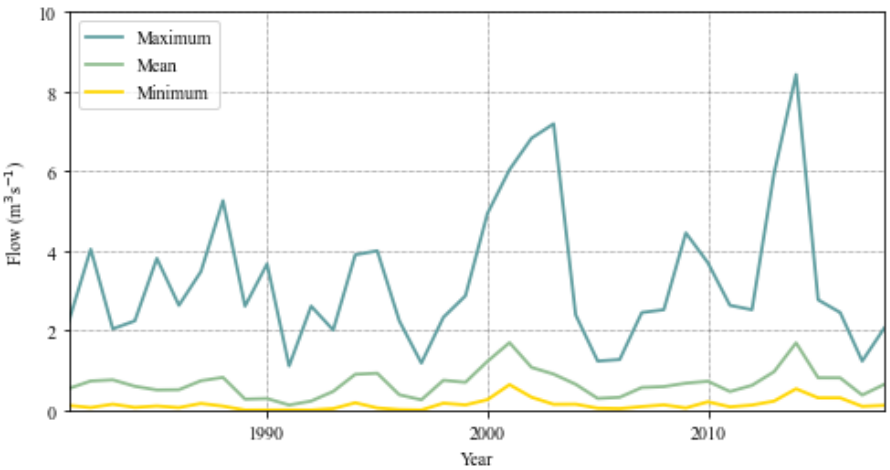


Fig. 6.12 Maximum, mean, & minimum flow by year for the River Darent at Hawley

In addition, the proxy variables for human influence, whilst useful in obtaining results that were, if not highly accurate, at least indicative of catchment behaviour for problematic rivers could be refined further. A better strategy might be to take a step back, to view some of the human influence on a catchment's behaviour as being internalised to some other process, in the same way that antecedent conditions were internalised, subordinate to a set of climatic variables and fixed catchment variables. The human influences within a catchment could be subordinate to the broader social, economic, and political factors; thus, a combination

of national and local metrics, which would include but not necessarily be limited to Gross Domestic Product per capita, population density, political leaning, the perceived level of corruption, the extent to which infrastructure is privatised or nationalised, could be highly correlated with and used as a means to approximate rates of water consumption, pollution and waste discharge (both legal and illegal), and the quality of water resources infrastructure. For example, recent political changes in the United Kingdom have resulted in the Environment Agency's budget being cut by more than 50% over the past decade, severely diminishing its capabilities to regulate effluent discharges and other negative interactions with water resources Greenwood (2020). Much of this transient data is collected and published through various international institutes, such as general economic data and perceived corruption by country (International Monetary Fund, 2022; Transparency International, 2022); ergo, there is both the potential and relative ease in data gathering to further drive predictive power and achieve even greater generalisation capability.

6.5 A Generalising Model Realised

The modelling approach described in this chapter was able to generalise across the United Kingdom with accuracy equivalent to those designed to work for and trained on a single catchment in many cases. With an expansion to the training set, perhaps both in presented permutations and size, then our approach has global potential whilst being computationally efficient and easy to transfer.

We also developed a methodology for describing variables that drive the model from a human influence perspective and that, as part of an ensemble modelling approach, can enable significantly improved predictions for catchments experiencing significant yet immeasurable human interaction. The philosophy presented here, in terms of encoding beliefs about system behaviour for parametric machine learning models likely has applications in adjacent fields, and perhaps even further.

Chapter 7

Sequential Process Models

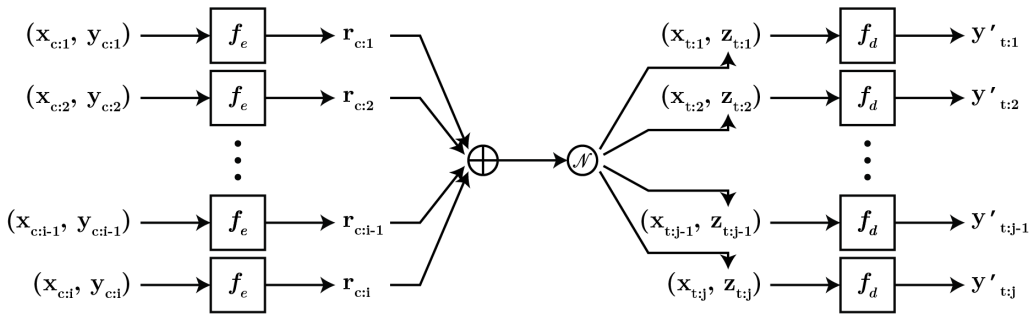
The Neural Process model, as discussed and used in Chapters 5 and 6, offers an alternative family of models that can combine the benefits of both deep neural networks and Gaussian Processes. Within that Neural Process lies flexibility around the choice of architecture and the standard detailed in an NP, whilst already shown to have significant potential, may not necessarily be the most optimal for sequence modelling purposes.

In this chapter, we pose a hypothesis based on the relative performance of the Neural Process when compared against the standard MLP, as follows: the performance of sequential neural network models might be improved by substituting them for the encoder and decoder networks and that the resulting model might offer further improvement over the base NP. In a sense, an analogy can be drawn between the kernel design in a GP being used to express prior beliefs about the relationship between a system's input and outputs and the architecture of the decoder and encoder networks here with regards to their suitability for a given problem. One might reasonably expect a range of different architectures being suitable for use within the Neural Process framework, based on the structure of the input data and the nature of the problem at hand.

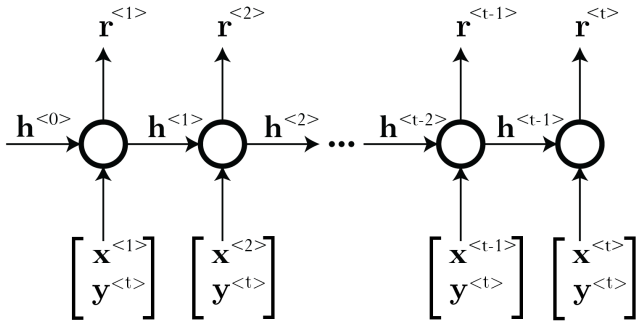
Thus, a pair of alterations are suggested that could be more applicable to the sorts of problems of focus here, those with history and where antecedent conditions are of significance. Thus, extensions to the Neural Process family are defined by the replacement architectures for its standard neural network, those of Recurrent Neural Networks and Temporal Convolutional Neural Networks, resulting in a Recurrent Process and a Temporal Process.

7.1 Recurrent Processes

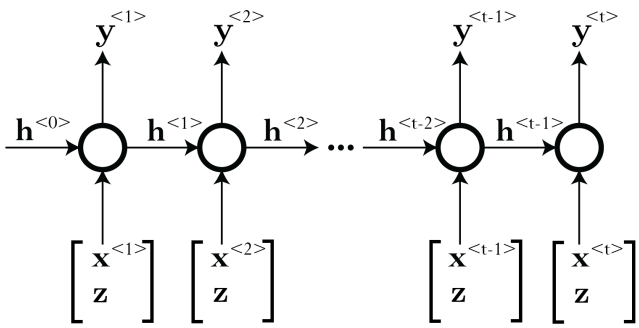
First, we examine the use of an RNN in place of both the encoder and decoder networks. Recall that the representation encoder maps the context pairs x_c and y_c to a representation



(a) Recurrent Process Architecture



(b) Recurrent Process Encoder



(c) Recurrent Process Decoder

Fig. 7.1 Architecture for the Recurrent Process model with encoder and decoder structure in terms of input and output

encoding r_c and the decoder maps the parameterisation and target pair, z and x_t, to a predicted target, y'_t with our metric of accuracy being the discrepancy between the mean predicted target, $\bar{y}'_t$, and the observed target value, y_t. Thus, for the Recurrent Process where capturing the hidden state between input steps is key, the encoder and decoder will be RNNs. By extension, and given the weights remain constant across time steps, the context observation and parameterisation may be constant across time steps for the encoder and decoder respectively, as shown in Figure 7.1, where the input and output context pairs, $\mathbf{x}_c^{<t>}$ and $\mathbf{y}_c^{<t>}$, are input to the encoder network to internalise hidden states, $\mathbf{h}_c^{<t>}$, and obtain the sequential encoding, $\mathbf{r}^{<t>}$. Similarly, the decoder network takes the input and parameterisation target pairs, $\mathbf{x}_t^{<t>}$ and $\mathbf{z}_t^{<t>}$, are input to the encoder network to internalise hidden states, $\mathbf{h}_t^{<t>}$, and obtain the target output, $\mathbf{y}^{<t>}$.

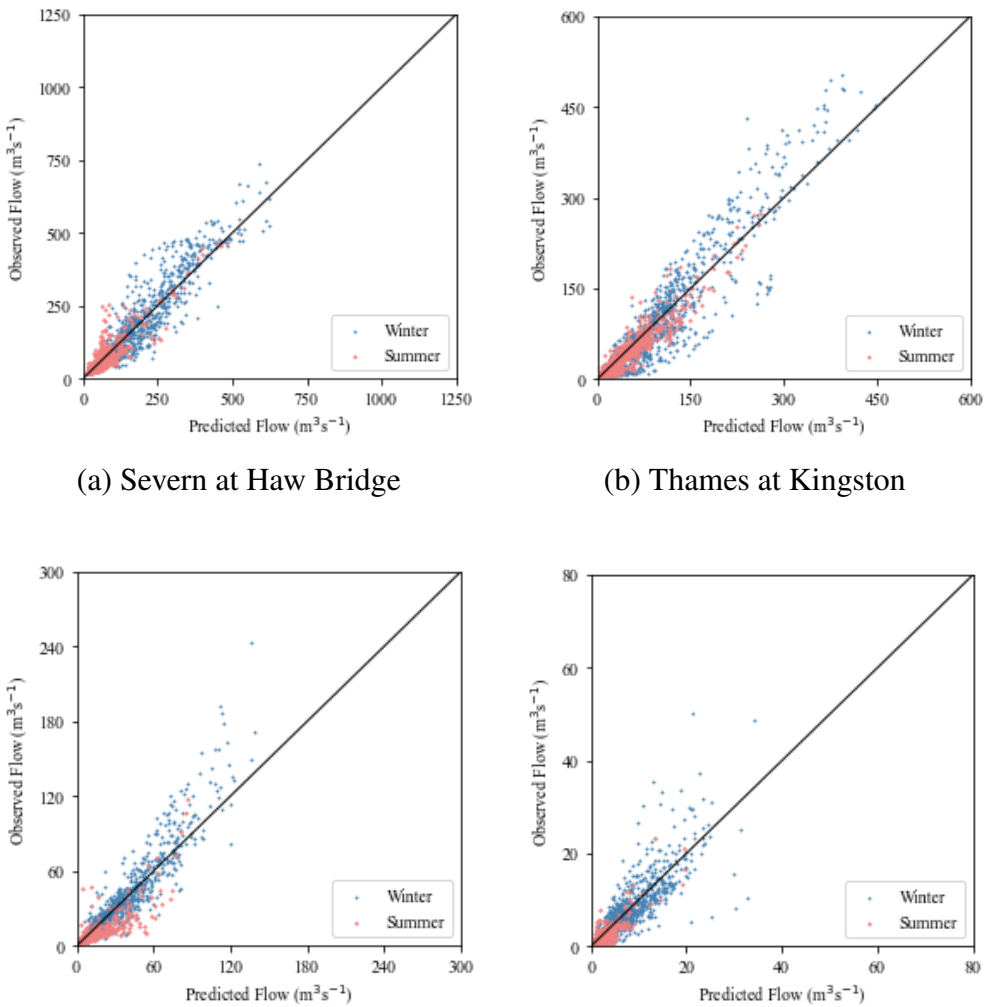
Our Recurrent Process model, as an extension of the Neural Process, inherits its core properties, such as scalability and the fact that it remains a conditional distribution over functions. We do note, however, that as RNNs must be evaluated and back-propagated sequentially, in terms of time steps, so too are the encoder and decoder networks, resulting in a higher computational complexity, which we found to be $\mathcal{O}(t \cdot (n+m))$.

7.1.1 Implementation

For this specific implementation of the Recurrent Process, the encoder and decoder architectures are taken from the RNN model used in Chapter 5, with 2 hidden layers and 128 hidden units for a 14 day sequence length. The training regime and train/test split are adopted from the Neural Process model used in Chapter 5. The results for this Recurrent Process are presented in Table 7.1 and Figures 7.2 & 7.3.

Table 7.1 RNP prediction performance by catchment

Model	Metric				
	NSE	NSE-RA	Q_1	Q_2	Q_3
Observations	-	-	11.7	29.9	79.5
Severn at Haw Bridge	0.887	0.850	37.8	73.0	145
Thames at Kingston	0.885	0.830	16.7	39.9	94.1
Avon at Bathford	0.868	0.832	4.55	10.1	27.0
Exe at Pixton	0.771	0.671	1.41	2.45	5.34
Average	0.852	0.796	-	-	-



(a) Severn at Haw Bridge

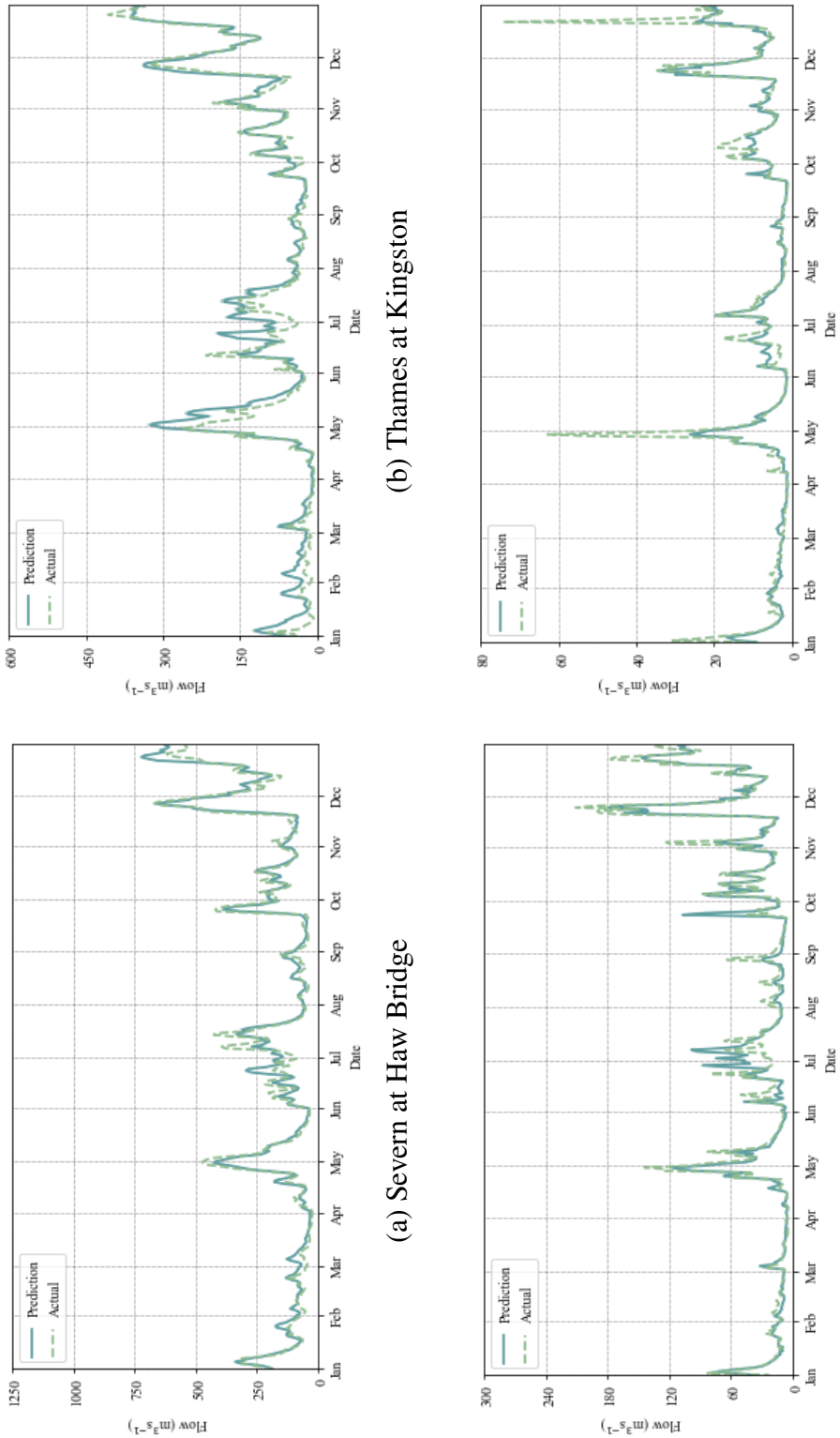
(b) Thames at Kingston

(c) Avon at Bathford

(d) Exe at Pixton

Fig. 7.2 Predictions against observations for the four test catchments using the Recurrent Process with $t = 14$

Based on the results, the performance of the Recurrent Neural Process exceeds that of all other methods for both NSE and NSE-RA on average, at 0.852 and 0.796 respectively; the predictive error around 2012 extremes for the River Exe persists, however. From the relative ease of implementation over the standard Neural Process with its increase in performance, we conclude that this model is highly suitable for sequential problems.



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 7.3 Comparative streamflow time series for the test set year, 2012, with predictions and observations using the Recurrent Process with $t = 14$

7.2 Temporal Processes

Intuitively, the improvement on the RNN using the Recurrent Process should translate to an improvement on the Temporal Convolutional Neural Network using a Temporal Process, being a Neural Process framework with a TCNN as the encoder and/or decoder.

Unlike the RNN, however, the $\mathbf{x}_c$ and $\mathbf{y}_c$ pairs and $\mathbf{x}_t$ and $\mathbf{z}$ pairs are not input into their respective networks together. Instead, we use the intrinsic structure of a TCNN to our advantage; the Convolutional part of the network is used to obtain a Temporal Encoding of the observations, such that we have a vector with a receptive field that encompasses the entire sequence, which we denote $TE(\mathbf{x}^{<1>}, \ldots, \mathbf{x}^{<t>}) = TE(\mathbf{x}_c)$. As with a TCNN, where a standard MLP is used after all convolutional layers, an MLP is used in the Temporal Process for $TE(\mathbf{x}_c)$ with its context pair, $\mathbf{y}_c$, for the encoder, $f_e\big(TE(\mathbf{x}_c), \mathbf{y}_c\big)$, and with the parameterisation, $\mathbf{z}$, for the decoder, $f_d\big(TE(\mathbf{x}_t), \mathbf{z}\big)$. This architecture is shown in Figure 7.4.

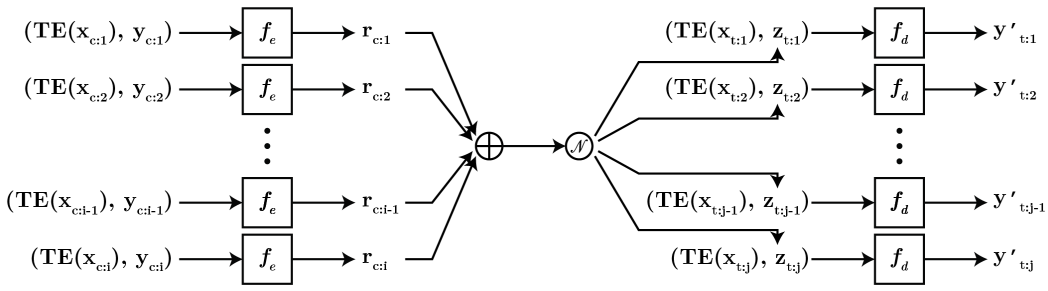
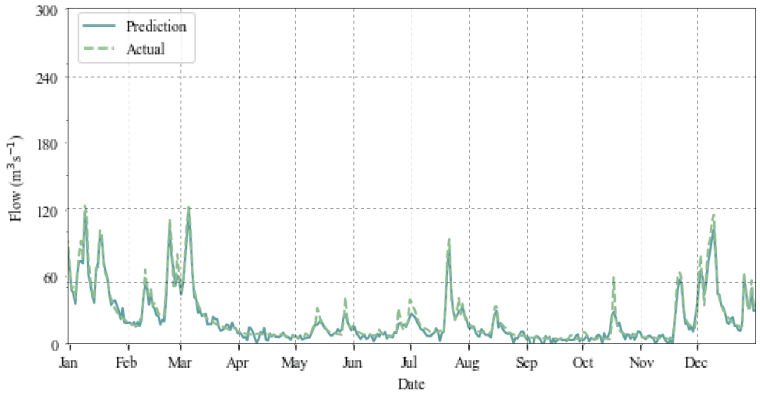


Fig. 7.4 Architecture for the Temporal Process model

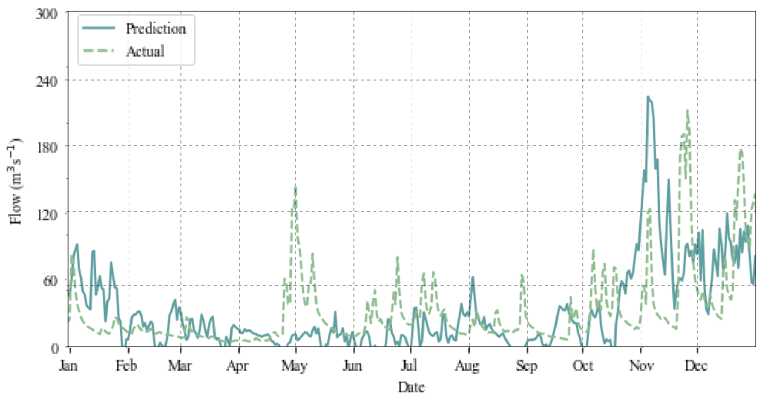
7.2.1 Implementation

As with the Recurrent Process, the TCNN structure from Chapter 5 is used to obtain the encoder and decoder Temporal Encodings, TE_e and TE_d respectively. The time dilation for both is set as 2 and the number of kernels was used as for the ordinary CNN, increasing as follows: $8 \rightarrow 16 \rightarrow 32 \rightarrow 64$.

Results for the Temporal Process, however, were less promising than for the Recurrent Process and the NSE and NSE-RA on the test set were 0.332 and 0.171, respectively, for the Avon at Bathford. Note that methods were used to prevent overfitting but there was still significant issue with overfitting, as shown in Figure 7.5, with the NSE and NSE-RA on the test set being 0.989 and 0.976, respectively.



(a) Performance of the Temporal Process on the training set



(b) Performance of the Temporal Process on the test set

Fig. 7.5 Temporal Process overfitting with the

Further attempts at regularisation failed to deliver significant improvement, with a combination of different levels of dropout, weight decay, and training stoppages being used. The best heavily regularised training regime resulting in less discrepancy between the performance on the training and test set but the gain in NSE accuracy was limited, improving to 0.350, with a sample of the test predictions shown in Figure 7.6.

With the Temporal Process proving difficult to train, alternative structures for the Temporal Encoders that feed into the model encoder and decoder, f_e and f_d, were investigated, to determine whether or not model structure was resulting in the overfitting and failure to generalise to the test set, with results shown in Table 7.2.

Varying the structure of the temporal encoding through the permutations listed in Table 7.2 had little discernible pattern on the performance of the model. If the Temporal Encoding

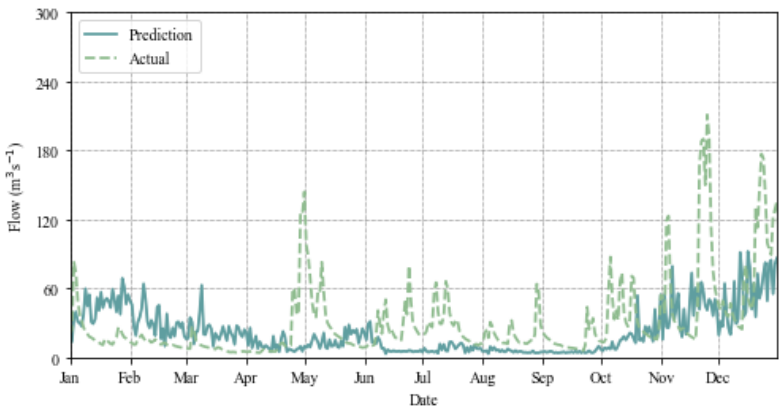


Fig. 7.6 Heavily regularised Temporal Process test predictions for the River Avon at Bathford in 2012

Table 7.2 Temporal Process prediction performance using different

Temporal Dilation	Temporal Encoder Kernels	NSE	NSE-RA
2	$8 \rightarrow 16 \rightarrow 32 \rightarrow 64$	0.350	0.182
	$8 \rightarrow 32 \rightarrow 128 \rightarrow 256$	0.347	0.170
	$8 \rightarrow 8 \rightarrow 8 \rightarrow 8$	0.294	0.095
	$4 \rightarrow 8 \rightarrow 16 \rightarrow 32$	0.329	0.165
3	$8 \rightarrow 16 \rightarrow 32$	0.327	0.173
	$8 \rightarrow 32 \rightarrow 128$	0.333	0.161
	$8 \rightarrow 8 \rightarrow 8$	0.321	0.125
	$4 \rightarrow 8 \rightarrow 16$	0.328	0.142

is removed and $x^{<i>} \ \forall \ 1 < i < t$ are inputs to f_e or f_d the model simplifies to the ordinary Neural Process, thus we infer that the cause of the performance is indeed the Temporal Encoding. However, we were unable to identify a method for approaching the optimal Temporal Encoding, if one exists. The strategy of selecting kernel numbers commonly adopted in CNN architectures, in terms of progression, as evidenced in Chapter 5 was unsuccessful and if an optimal Temporal Encoding exists that would achieve the desired performance, then a more elegant strategy for arriving at that optimum is required than what was followed here.

7.3 Discussion

Adapting the Neural Process model to include an RNN to form the Recurrent Process was a relatively simple step, with the RNN encoder and decoder architecture mirroring that of the RNN from Chapter 5, and the resulting model being more accurate than both the Neural Process and RNN models that inform it.

It was our belief that the hidden state captured by the RNN still being important to the nature of this problem that drove the adaptation of the standard Neural Process and that this is likely being captured within the encoder and decoder networks. Whilst the same concept underpins the Temporal Process, the performance gain was not realised, with the model proving difficult to fit and optimise.

We can likely conclude that further investigation into the Temporal Process model is required, in terms of its optimisation. Given that Conditional Neural Processes and their Convolutional counterpart are effective at learning with limited training sets (Garnelo et al., 2018; Gordon et al., 2019), the same should be expected with the Temporal Process model and augmenting or expanding the training set should not be expected to yield improved model performance. Instead, if the model can be improved, and based on the performance of the Recurrent Process and the prior art on Conditional Neural Processes then this is not an unreasonable assumption, the model architecture is likely where that improvement can be made. Whereas for the Recurrent Process the encoder and decoder architectures were assumed from the RNN with high performance, that was not the case with the TCNN; thus, again citing earlier commentary from Chapter 5, a more elegant model optimisation strategy is required, one which is able to efficiently handle the increased dimensionality of the optimisation space. That space necessarily includes both the encoder and decoder architecture along with the r and z parameters and training procedures.

7.4 Development Outcomes

In this chapter, we proposed new forms of the Neural Process model that might be more suited to environmental time series applications, of which hydrological phenomena form a subset. The Recurrent Process form is a high performance model that can be taken forward, underpinned by relative ease of implementation and low computational cost. The Temporal Process, however, currently requires further investigation.

As a secondary output to the functional model, we have also highlighted the flexibility of Neural Processes in being adapted towards specific applications. Depending on structure of data upon which the model is fitted, the encoder and decoder networks can either both of

separately be customised, for example to a geodesic data structure rather than linear time data structures.

Ultimately, we believe that both the Recurrent Process and Temporal Process models have the potential to deliver state of the art performance but the latter is more complex to train and implement, thus hindering its practical application. The former, however, is a highly suitable model for time series environmental science applications, such as streamflow prediction.

Chapter 8

Storm Prediction

To restate the objective for this chapter: the ideal outcome is for a machine learning model that can use circulation data to provide predictions of storm location, trajectory, and, subsequently, the hydrological influence or impact. The form that the latter takes is, simply, the localised increase in precipitation due to the storm that would then force a catchment hydrological model, as described in Chapters 4 to 7.

The distinguishing and eponymous feature of cyclones, their rotation, is of extreme importance; so much so that it will likely be integral to the development of algorithms here, as an input variable, and quantifying that rotation is necessary. Vorticity, $\vec{\omega}$, is the metric of choice here, defined as the curl of the velocity vector, $\vec{\Lambda}$, and is expressed as in Equation 8.1, with velocity components u, v, w in three dimensions x, y, z and corresponding unit vectors, $\hat{i}, \hat{j}, \hat{k}$.

$$\vec{\omega} = \nabla \times \vec{\Lambda} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ u & v & w \end{vmatrix} \tag{8.1}$$

$$= (\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}) \cdot \hat{i} + (\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}) \cdot \hat{j} + (\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}) \cdot \hat{k}$$

Relative vorticity, ζ, is the vertical component of vorticity, resulting in a simplification of vorticity, as expressed in Equation 8.2.

$$\zeta = \hat{k} \cdot \hat{\omega} = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{8.2}$$

8.1 Storm Identification

8.1.1 Basic Convolution

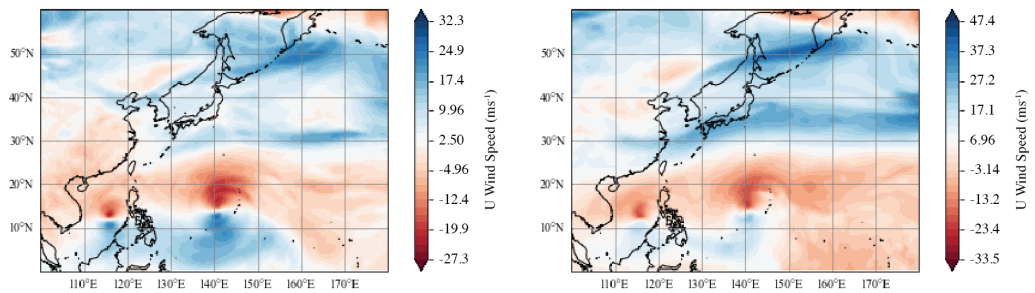
We begin solving our problem by first determining whether or not it is possible for an algorithm to accurately be able to predict if there is storm activity present within a domain. Our target output, y, is, therefore, a binary signal, such that $y \in [0,1]$, with the positive signal indicating storm presence.

The first model to be employed is one with which to examine the veracity of the approach; in other words, a simple convolutional neural network will be employed. The Convolutional Neural Network model, as described in Chapter 5, is largely the same in terms of implementation, though the following changes, or elaborations, should be noted: the input tensor is now 3-dimensional, rather than 2-dimensional, with dimensions equal to the longitudinal and latitudinal discretisation of the domain and the number of climatic variables of interest over the domain; although the output is still univariate, it now represents a probability. As such, a more appropriate loss function is the Binary Cross Entropy Loss (BCE), with formulation in Equation 8.3, where the Probability of a prediction, $P(y_i')$ being either a true or false value is compared against the actual observation y_i.

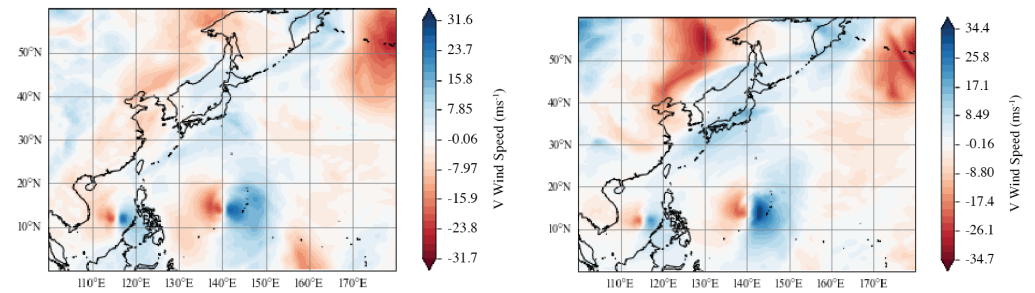
$$BCE = -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i \cdot ln\big(P(y_i')\big) + \big(1-y_i\big)\cdot ln\big(1-P(y_i')\big)\right\} \tag{8.3}$$

The dataset to be used for training and testing is a 15 year subset of the full record of storm data along with accompanying climatic variable data across the domain, with latitude ranging from 100°to 180°East and longitude ranging from 0°to 60°North. The feature variables for this experiment are the u and v components of windspeed, those in the plane parallel to the Earth's surface, relative vorticity, and relative humidity as observed at the 500hPa and 750hPa pressure levels.

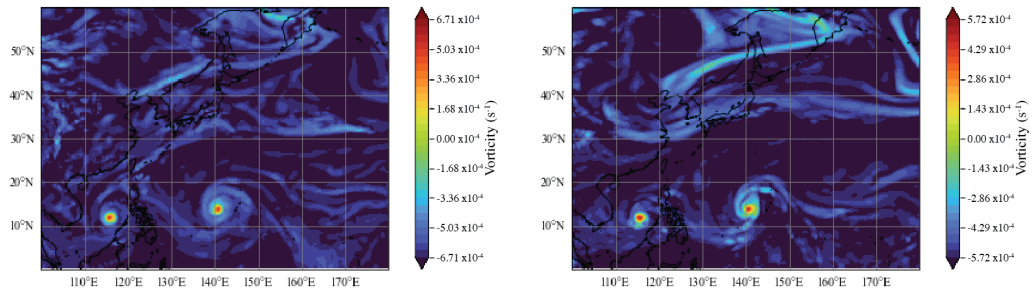
There are identifiable patterns within climatic variables associated with storm activity; for example, the two storms occurring simultaneously at 18:00 on the 10th of October 1979, Sarah & Tip, feature the high speed winds rotating about a core, as seen from the u and v components of wind speed, and are accompanied by spikes in relative vorticity towards the centre point of those wind speed patterns, as shown in Figure 8.1. Conversely, for the time step occurring approximately 15 days after their passing, these features are absent in the domain, as shown in Figure 8.2. Patterns in relative humidity also accompany the wind



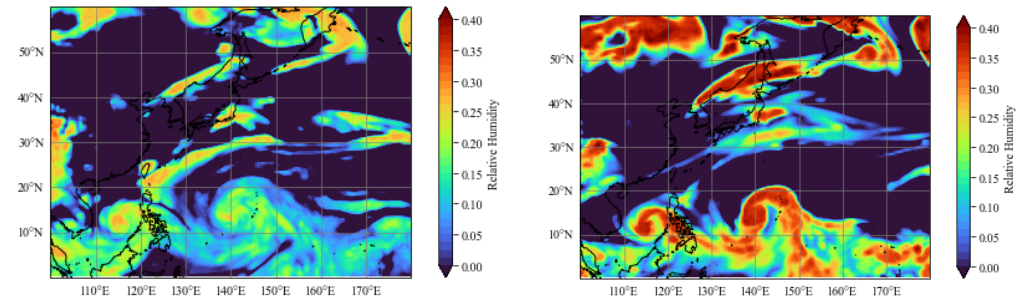
(a) U component of wind speed at 750hPa (left) and 500hPa (right)



(b) V component of wind speed at 750hPa (left) and 500hPa (right)

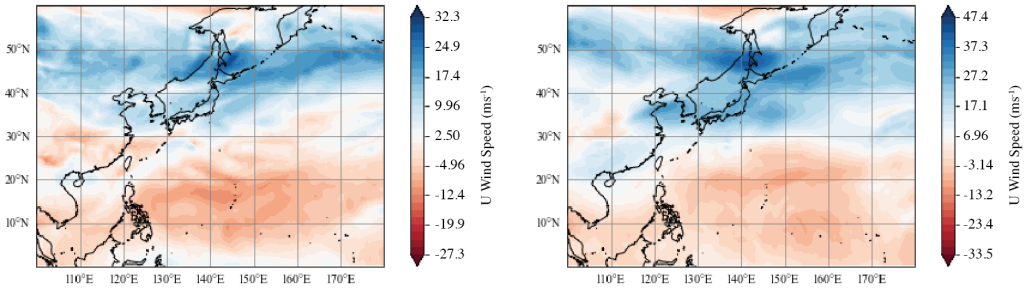


(c) Relative Vorticity at 750hPa (left) and 500hPa (right)

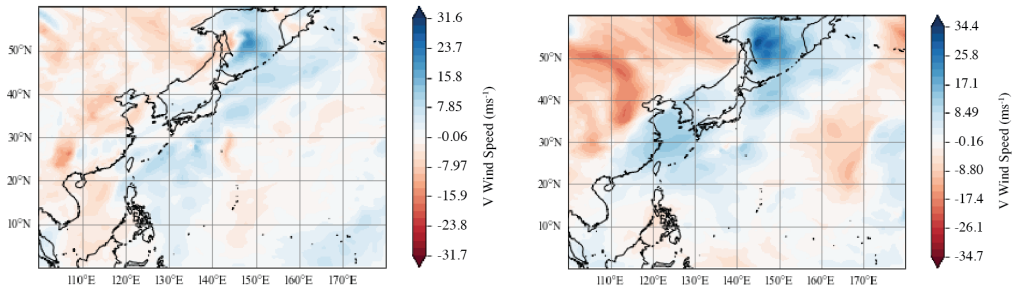


(d) Relative Humidity at 750hPa (left) and 500hPa (right)

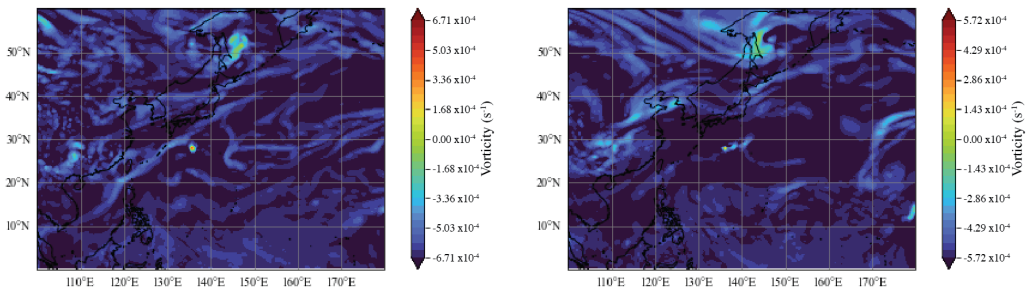
Fig. 8.1 Climatic variable patterns for the typhoons named Sarah & Tip, at 18:00 on the 10th of October 1979



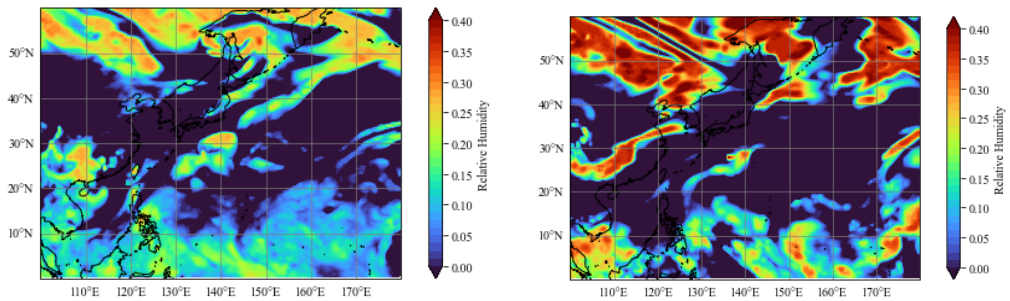
(a) U component of wind speed at 750hPa (left) and 500hPa (right)



(b) V component of wind speed at 750hPa (left) and 500hPa (right)



(c) Relative Vorticity at 750hPa (left) and 500hPa (right)



(d) Relative Humidity at 750hPa (left) and 500hPa (right)

Fig. 8.2 Climatic variable patterns approximately 15 days after the typhoons named Sarah & Tip have ended, at 12:00 on the 25th of October 1979

speed and vorticity patterns, in terms of spatial distribution, even if the absolute levels are unremarkable. Therefore, if these features are observable, applying a CNN to detect them is apt.

The architecture of the network is simple, featuring consecutive layers of 2D convolutional layers with 3x3 kernels, followed by batch normalisation, SiLU activation function, and a max pooling operation to downsample, as outlined in Table 8.1. This first pass at a CNN architecture for storm analysis is termed StormNet-01.

Table 8.1 StormNet-01 model architecture

Layer #	Layer type	Kernel #	Activation	Normalisation	Resampling
1	Conv 3x3	8	SiLU	BatchNorm	Max Pooling
2	Conv 3x3	16	SiLU	BatchNorm	Max Pooling
3	Conv 3x3	32	SiLU	BatchNorm	Max Pooling
4	Conv 3x3	64	SiLU	BatchNorm	Max Pooling
5	Conv 3x3	128	SiLU	BatchNorm	Max Pooling
6	Linear	-	SiLU	-	-
7	Linear	-	Sigmoid	-	-

In terms of model accuracy, the two metrics to be used for assessing performance are classification accuracy score, Ξ_a in Equation 8.4, and F1 score, Ξ_f in Equation 8.7. The former is more simplistic, simply being the proportion of classifications, either true positive, TP, or true negative, TN, that were correctly identified, whilst the latter if the harmonic mean of precision, Ξ_p, and recall, Ξ_r. Given that precision is the proportion of correctly predicted positives out of all predicted positives, Equation 8.5, and that recall is the proportion of correctly predicted positives out of all actual positives, Equation 8.6, F1 score essentially acts to give a form of average over false positive, FP, and false negative, FN, rates. Both of these metrics have the benefit of being on a scale where the maximal score is 1 and are therefore easily contrasted and compared.

$$\Xi_a = \frac{TP+TN}{TP+FP+TN+FN} \tag{8.4}$$

$$\Xi_p = \frac{TP}{TP+FN} \tag{8.5}$$

$$\Xi_r = \frac{TP}{TP+FP} \tag{8.6}$$

$$\Xi_f = \frac{2}{\frac{1}{\Xi_p} + \frac{1}{\Xi_r}} \tag{8.7}$$

For the basic convolutional neural network, as described above, Ξ_a was 0.877 and Ξ_f score 0.865; so, even with a rudimentary network, we have established a machine learning framework capable of identifying storm activity within the domain to a relatively high level of accuracy. The predominant area where StormNET-01 seems to struggle is with storm activity that has been labelled as such but which is less obvious; more concretely, this is the period of cyclone formation or dissipation. However, there are still storms within the test set that are somewhat obvious to manual inspection of the climatic variables and for which better performance could be expected.

8.1.2 Improved Architectures

The CNN architecture utilised in the initialisation of this problem could, as per prior commentary, be improved somewhat. Thus, additional architectures commonly applied to gridded data will be explored and compared. Architectures that have been used with success over recent years will be implemented, with the older first and more recent last, representing a progression in complexity whilst also examining the additional features added and the number of variables. After all, this thesis is still concerned with efficiency and if a minor performance increase comes at the cost of massively increased computational drain and complexity, then one might question whether or not the increased number of parameters is justifiable.

All of the architectures investigated will first be described, in terms of the setup of layer blocks along with any techniques or advances specific to that network. Comparative results are supplied in the ensuing and appropriately named section, Section 8.1.3.

AlexNet

The AlexNet algorithm was devised in 2012 and achieved, at the time of its implementation, superior performance on image classification tasks (Krizhevsky et al., 2012); given that this preliminary task is also a classification task, the use of an AlexNet derivative is appropriate.

The difference between the preliminary network and this derivative, dubbed StormNet-AN, is scale, in terms of the number of kernels and, hence, parameters within the network. The architecture for StormNet-AN is outlined in Table 8.2.

Table 8.2 StormNet-AN model architecture

Layer #	Layer type	Kernel #	Activation	Normalisation	Resampling
1	Conv 11x11	64	SiLU	BatchNorm	Max Pooling
2	Conv 5x5	192	SiLU	BatchNorm	Max Pooling
3	Conv 3x3	384	SiLU	BatchNorm	Max Pooling
4	Conv 3x3	256	SiLU	BatchNorm	Max Pooling
5	Conv 3x3	256	SiLU	BatchNorm	Max Pooling
6	-	-	-	-	Average Pooling
7	Linear	-	SiLU	-	-
8	Linear	-	SiLU	-	-
9	Linear	-	Sigmoid	-	-

Obviously, the input tensor from the previous section is not directly compatible with the input specified for AlexNet, being 224x224x3, but the fix is simple, adjusting the number of nodes in the first of the fully connected layers. The result is that our derivative network has significantly fewer parameters within those fully connected layers, resulting in a total of approximately 14.5 million parameters, as opposed to approximately 60 million in the original AlexNet. Another deviation made from the original is batch normalisation, which was introduced after AlexNet and, as the name might suggest, adds normalisation to the parameter layers to improve training and convergence by reducing the risk of exploding or vanishing gradients (Ioffe and Szegedy, 2015), certainly useful in the context of CNNs with millions of parameters to optimise.

VGGNet

AlexNet offered a substantial increase in the scale of the network by increasing the number of kernels within the layers; the network developed by the Visual Geometry Group at the University of Oxford, on the other hand, increases the number of layers over AlexNet in the VGGNet model (Simonyan and Zisserman, 2015). Increasing the scale, in terms of the number of hidden layers and, thus, parameters has been shown to improve the representation learned through feature maps and overall accuracy when using a CNN for image recognition

problems (Goodfellow et al., 2014), in contrast to the result shown in Chapter 4 for a standard ANN in fitting a regression function.

Ergo, we would, perhaps, expect an improved result over StormNet-AN from our derivative of VGGNet, StormNet-VGG. The StormNet-VGG model has approximately 101 million parameters, reflecting the increased depth of the network, with the structure outlined in Table 8.3. Again, batch normalisation, a deviation from the original VGGNet model, is included to improve training.

Table 8.3 StormNet-VGG model architecture

Layer #	Layer type	Kernel #	Activation	Normalisation	Resampling
1	Conv 3x3	64	SiLU	BatchNorm	-
2	Conv 3x3	64	SiLU	BatchNorm	Max Pooling
3	Conv 3x3	128	SiLU	BatchNorm	-
4	Conv 3x3	128	SiLU	BatchNorm	Max Pooling
5	Conv 3x3	256	SiLU	BatchNorm	-
6	Conv 3x3	256	SiLU	BatchNorm	-
7	Conv 3x3	256	SiLU	BatchNorm	Max Pooling
8	Conv 3x3	512	SiLU	BatchNorm	-
9	Conv 3x3	512	SiLU	BatchNorm	-
10	Conv 3x3	512	SiLU	BatchNorm	Max Pooling
11	Conv 3x3	512	SiLU	BatchNorm	-
12	Conv 3x3	512	SiLU	BatchNorm	-
13	Conv 3x3	512	SiLU	BatchNorm	Max Pooling
14	Linear	-	SiLU	-	-
15	Linear	-	SiLU	-	-
16	Linear	-	Sigmoid	-	-

ResNet

So far, the only modifications made have been those that adjust the size of the model; the previous three architectures described, the basic CNN implemented at first followed by AlexNet and VGGNet derivations, present little more than a progression in the number of parameters and depth. Whilst this likely improves the complexity of the features that the model is able to learn, in practice increasing the number of layers can result in decreasing model accuracy with the problem potentially being indicative of difficulties in the development of appropriate mappings between multiple nonlinear layers (He and Sun, 2015).

The ResNet architecture (He et al., 2016), however, was devised in order to improve the training of deep CNNs through residuals being passed through a connection between stacks of layers such that the desired mapping after this residual "block", $R(\mathbf{x})$, is the sum of the nonlinear mapping applied to the input, $F(\mathbf{x})$, and the input itself, $\mathbf{x}$, as shown in Figure 8.3.

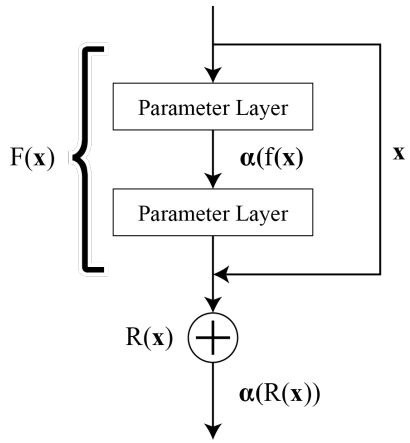


Fig. 8.3 ResNet module schematic

The ResNet50 model has 50 parameter layers, most of which sit within residual blocks featuring the skip connection. The StormNet-R50 model is constructed in very similar fashion, with the architecture outlined in Table 8.4; as the ResNet50 model already makes use of the amendments that had to be introduced for the AlexNet and VGGNet derived models, the modifications here are limited to activation function and output.

SEResNet

An elaboration to the ResNet framework is that of squeeze and excitation (Hu et al., 2018), implemented to enhance learning of convolutional features through the explicit modelling of channel interdependencies; within a squeeze-excitation (SE) module applied to a tensor, a single variable is obtained for each channel before being transformed by learned parameters, requiring optimisation, and then used to scale each of the respective channels that they represent. Elaborating further, the squeeze operation is often an averaging operation, noting that alternative functions could be used, applied channel-wise, whereby the output statistic, $\mathbf{z} \in \mathbb{R}^c$ for c channels, for a $d_h \times d_w \times d_c$ tensor, is generated per channel, z_c, as follows in Equation 8.8:

Table 8.4 StormNet-R50 model architecture

Layer #	Layer type & Kernel #	Activation	Normalisation	Resampling
1	Conv 7x7, 64	SiLU	BatchNorm	Max Pooling
2-4	[Conv 1x1, 64 Conv 3x3, 64 Conv 1x1, 256] x3	SiLU	BatchNorm	-
5-8	[Conv 1x1, 128 Conv 3x3, 128 Conv 1x1, 512] x4	SiLU	BatchNorm	-
9-14	[Conv 1x1, 256 Conv 3x3, 256 Conv 1x1, 1024] x6	SiLU	BatchNorm	-
15-17	[Conv 1x1, 512 Conv 3x3, 512 Conv 1x1, 2048] x3	SiLU	BatchNorm	-
19	-	-	-	Average Pooling
20	Linear	Sigmoid	-	-

$$z_c = \frac{1}{d_h \times d_w} \sum_{j=1}^{w} \sum_{i=1}^{h} x_{ij} \tag{8.8}$$

The excitation operation transforms the squeeze output statistics via learned parameters, Φ_1 and Φ_2, and activation functions, α, as per Equation 8.9, to generate a channel-wise scaling vector, κ. This is then applied to the original input tensor, $\mathbf{x}$, to obtain the output y_κ, in Equation 8.10.

$$\kappa = \alpha_{sigmoid}\big(\Phi_2 \alpha(\Phi_1 \mathbf{z})\big) \tag{8.9}$$

$$y_\kappa = \kappa \cdot \mathbf{x} \tag{8.10}$$

Together, the two operations form an SE block that can be used to modify the ResNet block, as in Figure 8.4, and act to promote features at different points within the network. The benefit of this is more intuitive if considered alongside, say, certain input variables or feature maps, with some being more informative than others. The derived architecture for our problem, StormNet-S50, is based on the SE block modification to StormNet-R50 and outlined in Table 8.5.

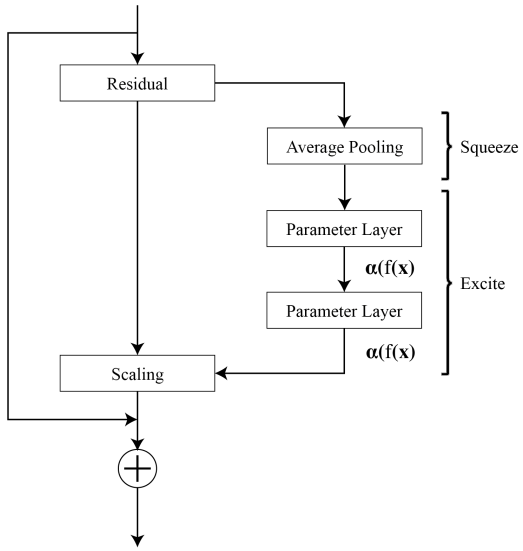


Fig. 8.4 ResNet Squeeze-Excitation module schematic

EffNet

The final architecture described in this section is that of the EfficientNet (Tan and Le, 2019), a framework for automatically increasing model scaling, across width, depth, and resolution. One of the key building blocks is the use of separable, depth-wise convolutions to maximise parameter efficiency; this applies $k \times k \times 1$ kernels over the input by channel before combining the resulting n feature maps and then applying a $1 \times 1 \times n$ kernel to form a 1 channel output tensor. Furthermore, the residual block is "inverted", moving the skip connection to connect the narrow channel layers rather than wide channel layers. In practice, this involves expanding the input tensor with a large number of 1×1 kernels to obtain a high

Table 8.5 StormNet-S50 model architecture

Layer #	Layer type & Kernel #	Activation	Normalisation	Resampling
1	Conv 7x7, 64	SiLU	BatchNorm	Max Pooling
2-4	[Conv 1x1, 64 Conv 3x3, 64 Conv 1x1, 256] x3	SiLU	BatchNorm	-
5-8	[Conv 1x1, 128 Conv 3x3, 128 Conv 1x1, 512] x4	SiLU	BatchNorm	-
9-14	[Conv 1x1, 256 Conv 3x3, 256 Conv 1x1, 1024] x6	SiLU	BatchNorm	-
15-17	[Conv 1x1, 512 Conv 3x3, 512 Conv 1x1, 2048] x3	SiLU	BatchNorm	-
19	-	-	-	Average Pooling
20	Linear	Sigmoid	-	-

channel tensor prior to performing the depthwise convolution and then squeezing back down using a small number of 1×1 kernels; this will be referred to as the MBConv-*er*, where *er* refers to the expansion ratio, and was adopted from the MobileNet architecture (Sandler et al., 2018). Thus, for the base EffNet model, StormNet-EB0, the architecture is as in Table 8.6.

The scaling of the architecture is done through a parameter, κ, to scale depth, width, and resolution uniformly. One issue that should be noted is how the network scales with regards to the number of input channels; the original EffNet models were all tested on an input comprising the standard RGB input channels. Our problem, although treated in the same way as a colour image, is actually one of considerably more channels. Therefore, the EffNet scaling for our version, StormNet-EB1, results in a model that has approximately 60.0 million parameters instead of the 27.2 million that it would have for an input tensor with only 3 channels.

Table 8.6 StormNet-EB0 model architecture

Layer #	Layer type	Kernel #	Activation	Normalisation	Resampling
1	Conv 3x3	32	SiLU	BatchNorm	-
2	MBConv-1 3x3	16	SiLU	BatchNorm	-
3	MBConv-6 3x3	24	SiLU	BatchNorm	-
4	MBConv-6 3x3	40	SiLU	BatchNorm	-
5	MBConv-6 3x3	80	SiLU	BatchNorm	-
6	MBConv-6 3x3	112	SiLU	BatchNorm	-
7	MBConv-6 3x3	192	SiLU	BatchNorm	-
8	MBConv-6 3x3	320	SiLU	BatchNorm	-
9	Conv 1x1	1280	-	-	Average Pooling
10	Linear	-	Sigmoid	-	-

8.1.3 Comparison

The comparison must, of course, be a fair one, so the training set and test set are the same across all experiments with multiple runs across permutations of learning hyperparameters, specifically learning rate and regularisation, for each model. Training is stopped when the models appear to have converged for test and validation error.

The results for Accuracy and F1 scores are shown for all models in Table 8.7 and are roughly in line with expectations in that, as the model architecture is scaled up and refined, the performance of the model improves. Much of that improvement is for the situations previously identified, in the neighbourhood of storm formation and dissipation, cyclogenesis and cyclolysis respectively.

Table 8.7 Storm activity prediction performance by model type

Model	Metric		
	Accuracy	F1 Score	Parameters (10^6)
StormNet-01	0.877	0.865	0.16
StormNet-AN	0.891	0.888	14.5
StormNet-VGG	0.890	0.888	101
StormNet-R50	0.909	0.908	47.1
StormNet-S50	0.910	0.908	51.6
StormNet-EB1	0.913	0.910	60.0

One might perhaps question the merit of the increase in accuracy when compared against the number of parameters, which necessarily increase the computational demand of running the model and, consequentially, training time. However, though the performance gain for this task is small as the number of parameters is increased, the computational power required for running these models is still well within the bounds of that made available for this work and was less of a constraint than the data loading strategy itself, which was demanding in terms of CPU rather than GPU load and is discussed further at the end of this chapter. Furthermore, the task of binary classification is simpler than the ensuing tasks to which the same methodology is implied; thus, the relative performance between model types may diverge. Rather than continuing with the comparison between model types, only the most accurate one, StormNet-EB1, will be used from hereon out.

8.2 Path Tracking

With the ability to identify storm presence established and predictions being of a sufficiently high level of accuracy, the problem now shifts to one of being able to accurately pinpoint the centre of a storm for the purposes of generating a typhoon track. Our target, therefore, becomes a coordinate pair, $[C_x, C_y]$, corresponding to the longitude and latitude of a storm's epicentre.

The highest performing algorithm, StormNet-EB1, from the previous experiment is retained in terms of the overall architecture and the approximate number of parameters with the output now being of higher dimensionality than before as continuous longitude and latitude variables to be optimised through MSE; at some points in time, there are multiple storms present within the domain, so the number of target pairwise coordinates is increased to allow for the prediction of the location of the ith storm, $[C_{x_i}, C_{y_i}]$. In practice, there is no need to allow for $i > 5$, given the highest number of simultaneously occurring storms in the training set and the equivalent number in the test set.

General performance for a storm track across the domain suggests that there is sound rationale in applying machine learning for locating storm activity; for example, the algorithm is able to locate the storm x, as seen in Figures 8.5 and 8.6, across most time steps with the predicted location being within the storm's radius. However, two issues arise: the first being that for $i > 2$, predictive capability diminishes, due to the few instances where there are more than 2 storms present simultaneously; and the second is that for cyclogensis and cyclolysis, the model struggles to pinpoint the location. For the latter, given the global location in terms of longitude, if we have a timestep labelled as positive storm activity and provide a coordinate pair for either cyclogensis or cyclolysis and the model fails to locate is accurately, then the

resulting error term can be disproportionately high. As a result, the RMSE for the model when predicting storm location for the test set is approximately 6.71°in latitude, 25.9°in longitude, even though the R^2 is 0.569 in latitude and 0.752 in longitude, whilst the average distance from the storm epicentre is 8.63°, approximately 900km.

In order to improve the predictive capability, transforms of the coordinate system were applied, through altering either the longitude and latitude scaling and the frame of reference to place the origin in the centre of the domain. However, neither significantly improved the error, with the resulting RMSE and R^2 values showing a less that 1% improvement. Alternately, removing cyclogensis and cyclolysis from the assessment metric offers significant improvement, cutting the longitudinal RMSE to 8.23°. We also note that although the RMSE for longitude is worse, the R^2 value is better than for latitude, due to the frame of reference. However, the error is still noticeably large; ergo, rather than focusing on the metric, we believe there to be a more fundamental flaw with the method.

Given that the training regimen and set size are as for the first task of identifying the presence of storm activity, the aspect likely the crux of performance is the manner in which the data has been labelled. The use of centre-point coordinates is likely ineffective in comparison to using a bounding box intersection-over-union (IoU) method. In order to learn multiple landmarks effectively, there would need to be appropriate representation for all sets of coordinate pairs; however, the number of timesteps with 3, 4, or even 5 storms represent ever smaller fractions of the dataset. Whilst those timesteps with higher numbers of storms are typically found when prior or subsequent storms during cyclogenesis and cyclolysis, the fact remains that there can still be some temporal overlap and the performance for the higher numbered coordinate pairs is lacking.

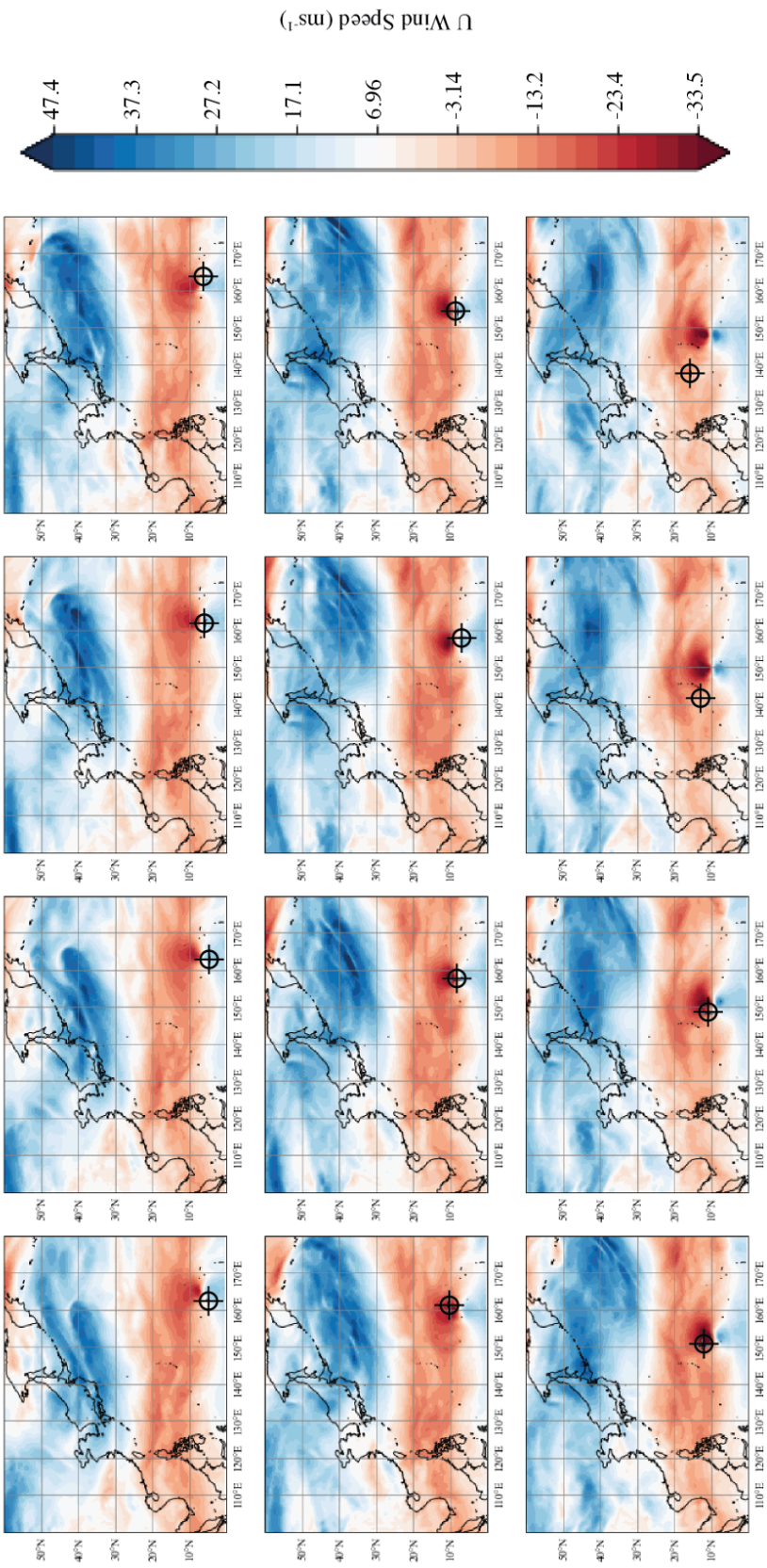


Fig. 8.5 Predicted location of storm , marked by a black cross, from 18:00 on the 10th of October 1979 to 18:00 on the 10th of October 1979, on u component of wind speed maps

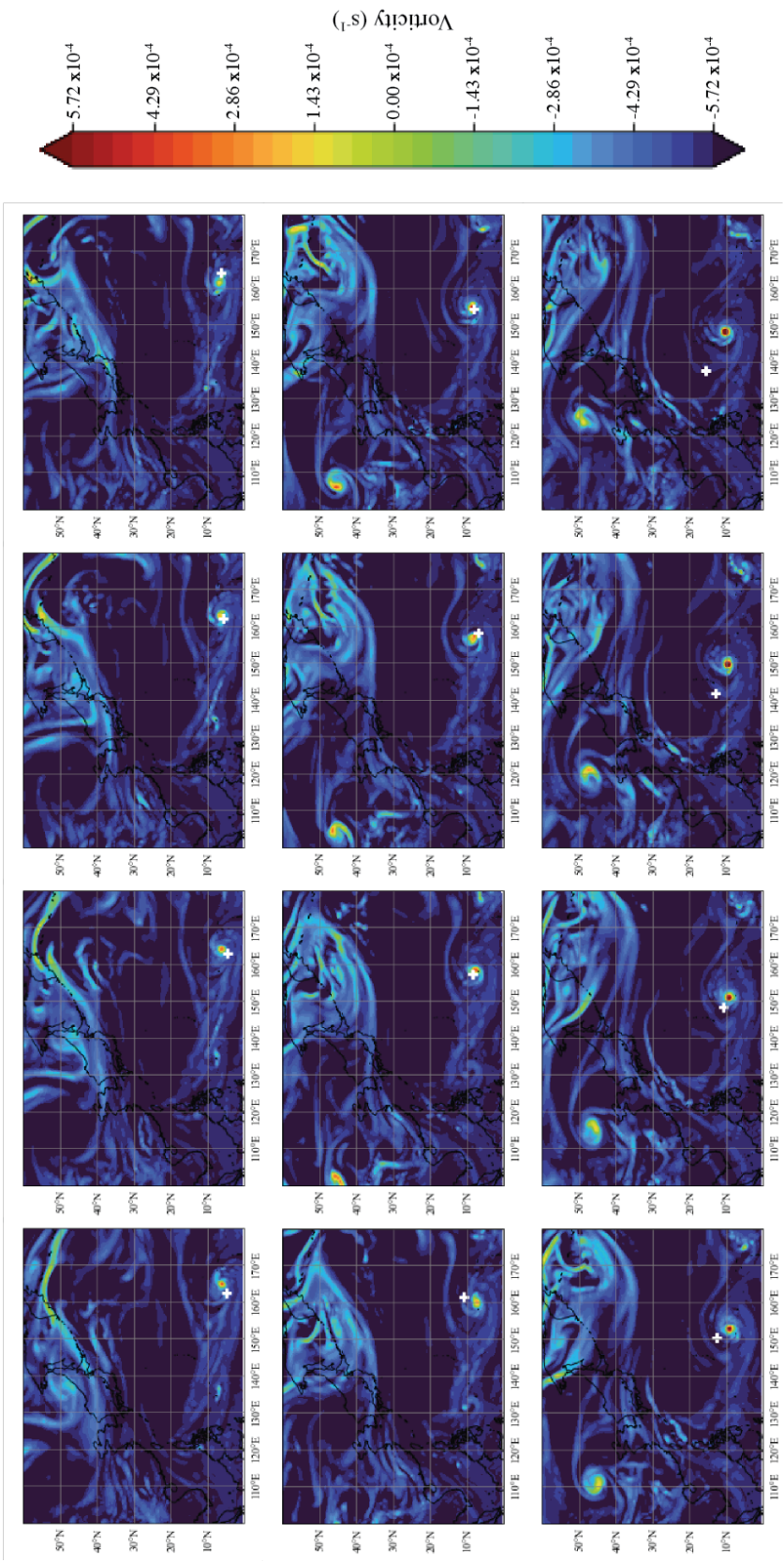


Fig. 8.6 Predicted location of storm , marked by a white cross, from 18:00 on the 10th of October 1979 to 18:00 on the 10th of October 1979, on relative vorticity maps

The benefit of using an IoU method is that it can be almost agnostic to the number of storms within the domain as, rather than trying to fit a preset number of coordinate pairs, the model would be systematically examining a subset of the domain for storm activity and identifying its presence, or lack thereof, within that subset. Whilst this would necessarily entail relabelling the dataset, as a subset within each tensor that corresponds to storm activity would have to be labelled as so rather than just a coordinate pair vector, another opportunity is presented to have the algorithm recognise radius of storm wind. Further splitting that new labelling into subcategories for cyclogenesis and cyclolysis would likely improve its robustness.

8.3 Precipitation Impact

Beyond the potential wind damage, the precipitative impact of storms is also potentially disruptive; major typhoons in the North West Pacific basin have caused significant damage and in some instances overwhelmed flooding defences (Hsu et al., 2015; Luu et al., 2021; Trošelj and Lee, 2021). Therefore, being able to accurately assess the increased precipitation attributable to a storm is an important part of assessing their overall impact and that of flood risk in general.

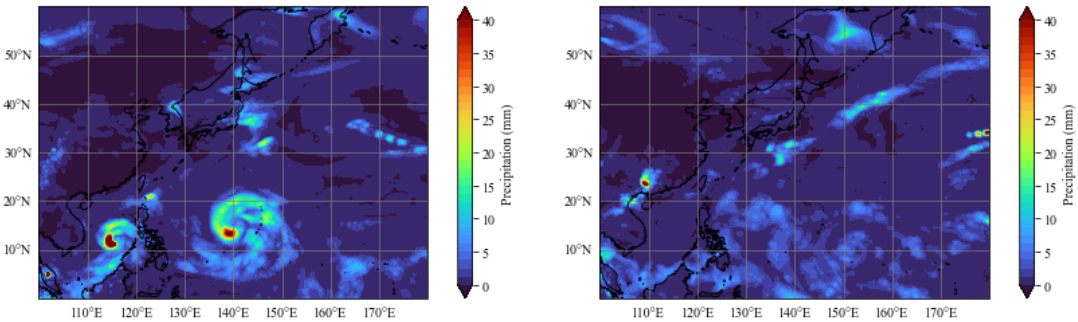
The target is the relative increase in precipitation, Ψ_r, being the difference between that which is caused by the storms, Ψ_s, over the domain average, $\bar{\psi}$ at a given 6-hourly interval, t, as expressed in Equation 8.11. The average being calculated as the cumulative rainfall over the latitude and longitude extents, x and y.

$$\Psi_r = \Psi_s - \frac{1}{xy} \int_{t_i}^{t_{i+1}} \int^{y} \int^{x} \psi(x,y,t)\, dx\, dy\, dt \tag{8.11}$$

No correlation was found between the average domain precipitation and the presence of storm activity within the domain; given a large enough domain, as indeed the North West Pacific Basin is, then cyclonic activity, as occurring over a fraction of that domain, should have relatively little impact in the overall transport of water within that domain, especially when compared to seasonal flux.

Conversely, when looking at the distribution of precipitation across the domain, relative to the average at that point in time, a clear pattern emerges. For all storms there is a significant increase in precipitation intensity within the locality of the storm, with the coordinates of

the maximum coinciding almost exactly with the maximum vorticity; for example, when referring back to the storms presented in Figure 8.1, Typhoons Sarah and Tip, the maximum vorticity and the maximum precipitation, presented in Figure 8.7, intensity occur at the same point. For the same counterpoint period as used before, there is no clear pattern in precipitation, as was the case with the other climatic variables, and no severe increase in specific localities.



(a) Precipitation during typhoons Sarah & Tip

(b) Precipitation after the typhoons have ended

Fig. 8.7 Total precipitation per six hour interval patterns during typhoons Sarah & Tip, at 18:00 on the 10th of October 1979, and approximately 15 days after the typhoons have ended, at 12:00 on the 25th of October 1979

Therefore, we posit that, given it has been possible to identify and locate storm activity, it may be entirely possible to link the relative increase in precipitation to that storm activity. The frameworks above are further extended once more to enable determination of the precipitative impact of a tropical storm based on its size and intensity. The size and intensity of the typhoon need to be defined in this context. There are measures for quantifying a typhoon, such as Power Dissipation Index, Typhoon Class, and Radius of Storm Wind, as provided in the database used for this study, but rather than using derived measures, all of which are essentially functions of wind speed, the raw input of these will be used, being the same dataset as for the previous problems.

With the model already defined and the input space the same as the previous iterations, the only modification is the change in target variable. Again, model weights are stored from previous iterations to minimise training time, with the exception of the final layer of the network. After fitting the model, the predictive capability of the model is shown in terms of

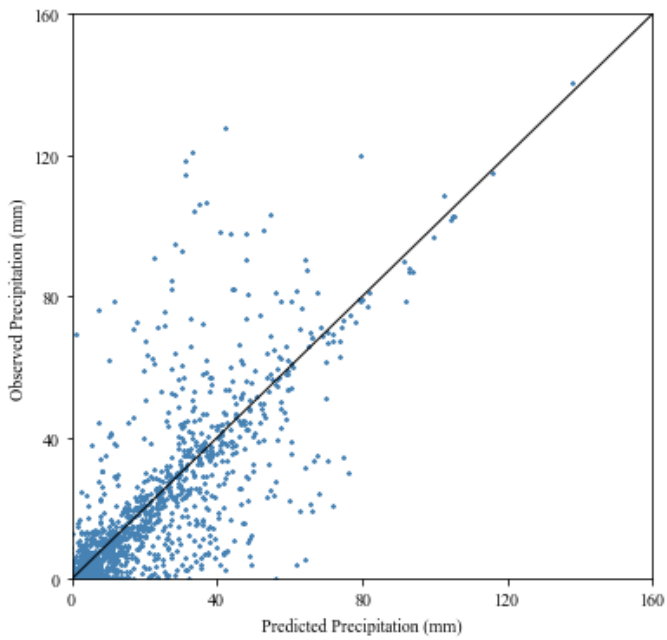


Fig. 8.8 Predictions against observations for storm-related precipitation, taken as the total precipitation over a six-hour interval, above the domain average at that same six-hourly interval

predictions versus observations in Figure 8.8 and through the RMSE of 11.1mm and and R^2 score of 0.652.

Thus, we enable the use of a single algorithm able to identify storm activity, track the location of said activity, and also predict the precipitative impact associated with it.

8.4 Future Development

The work presented here has highlighted the potential in the identification and classification of storms, including approaches to locating a typhoon within a domain, and predicting its precipitative impact for the purposes of projecting downstream flood risk and damage assessment. Whilst some aspects of the modelling techniques utilised here have been highly successful, others have been less so and are in need of refinement or improvement before the approach detailed here could be deployed for real world application.

One of the main deficiencies that has likely hindered the development of more accurate results is the amount of data that was made available to the algorithm in training. This has not arisen out of a lack of available data on the whole but rather through the sequencing algorithm and data loading strategy being inefficient, utilising the entirety of the available

CPU and GPU power available. Ideally, the amount of data, and therefore distinct examples, for the model to learn from should be increased to the entire record to improve the quality of predictions. With most forms of neural networks, convolutional being no exception, more data with greater variety of examples in the training set improves performance. Alternatively, few-shot learning techniques could be employed, such as the Conditional Convolutional Neural Process, noting that these models have other benefits in being probabilistic, achieving state of the art performance, and modelling translation equivariance (Gordon et al., 2019).

As was mentioned with the path tracking section of this chapter, extensive dataset relabelling in order to enable IoU prediction of storm could deliver significant gains in performance. Whilst this task would be conducted manually and therefore be time intensive, it would perhaps be a complementary activity to addressing the data loading strategy, for combining the datasets in input-output target pairs.

One area of interest is in extending prediction of the storm location and its impact through an additional t time-steps; the deep learning probabilistic IceNet model, presented by Andersson et al., 2021, utilises a similar input dataset, in terms of variables at various pressure levels, along with sea conditions in a U-Net architecture to predict how sea ice will evolve in future time-steps with performance (Andersson et al., 2021). Due to this similarity, between the input space and the target outcomes, a similar model structure could be utilised to make that forward projection. The input space, however, will need to be expanded to include variables that have an impact on the formation and dissipation of storms, namely sea surface temperature and wind speeds at additional and lower pressure levels, e.g. 250hPa, as this would provide a richer picture of wind shear, potentially leading to the dissipation of cyclones. Topography would also be a useful input, due to the potential interactions with cyclones.

8.5 A Positive Trend

What has been achieved, in spite of the obvious limitations of the method as implemented in this thesis, is to showcase how a machine learning algorithm can be used to extract useful information from gridded circulation data in terms of the impact of extreme phenomena. Should the data labelling strategy be revisited and improved, as per the suggestions in the preceding section, then we would expect that the value of this approach would deliver upon its promise, tracking the location of a storm and localising its hydrological impact. Thus, we consider the modelling approach itself a viable output, subject to more appropriate data labelling.

Chapter 9

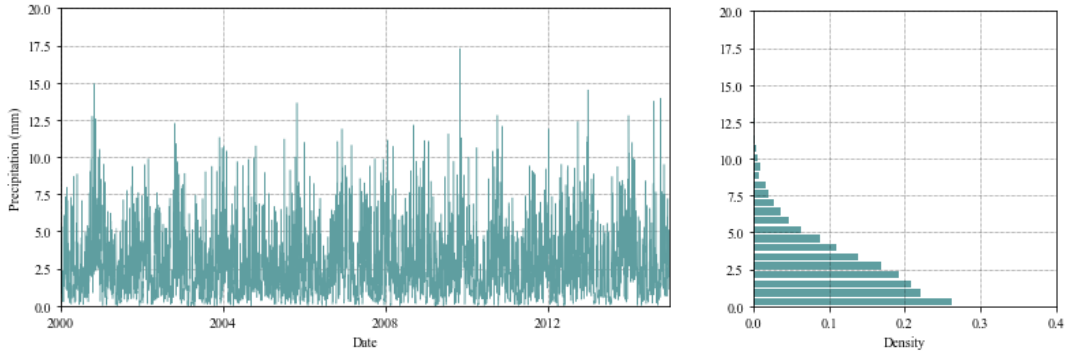
Impact Modelling

In this final chapter, the task now turns to real-world, integrated application, the use of these models in conjunction with CMIP6 data for the generation of potential hydrological scenarios under the different emissions scenarios. First, a comparison between the raw output of a GCM and observations, along with a discussion of the need for correcting systemic model biases, is presented before data from a CMIP6 constituent model is utilised to generate hypothetical scenarios with discussion on the implication and overall suitability.

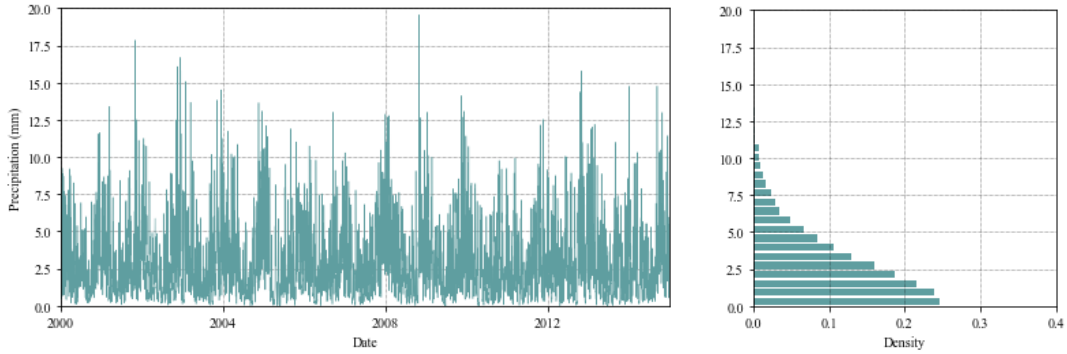
9.1 General Circulation Model Data

In accordance with the data strategy outlined in Chapter 3, a single model approach is being explored first, using data from the MPI-ESM for the SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5 scenarios, extracted over the subdomain of relevance over the time period 2014 to 2100 along with the historical record from 1979 to 2014. To keep the presentation of data succinct, the two most important climatic signals for this problem, precipitation and temperature, are shown in Figures 9.1 and 9.2, respectively, for a subset of the time series, 2000 to 2014, along with the model error between the observations and the hindcasts. Note that the model output for precipitation is actually precipitative flux, with units of $\mathrm{kgm^2s^{-1}}$ rather than mm, and is corrected by multiplying by requisite time period over the density, to obtain precipitation depth.

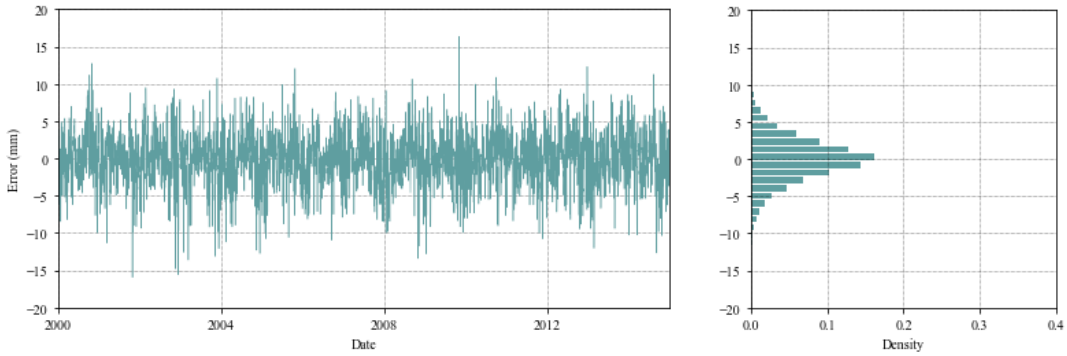
From the error signal between observations and the hindcasts, that there is systematic model bias is transparent; extant for all climatic variable signals, with sometimes significant error, a strategy for correcting it is required. Given our interest in the more direct use of GCM data in forcing machine learning models, we will initially consider the use of model output statistics. Thus, an investigation of Quantile Mapping is provided with a view to elucidating



(a) Time series and associated histogram of observations

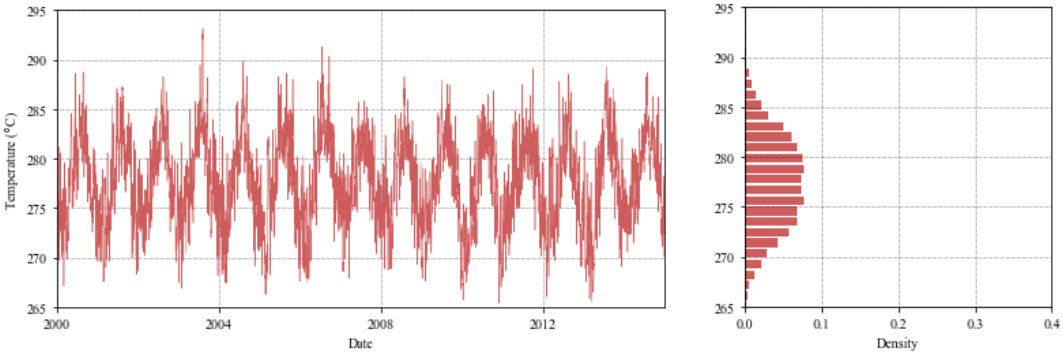


(b) Time series and associated histogram of hindcasts

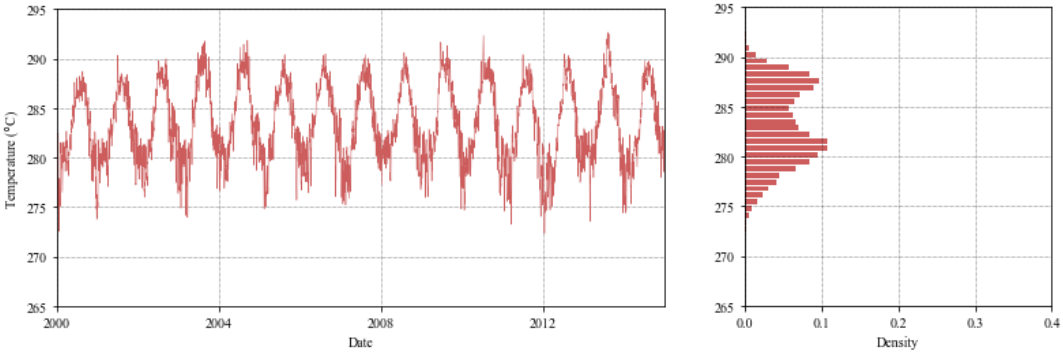


(c) Time series and associated histogram of the observation-hindcast error

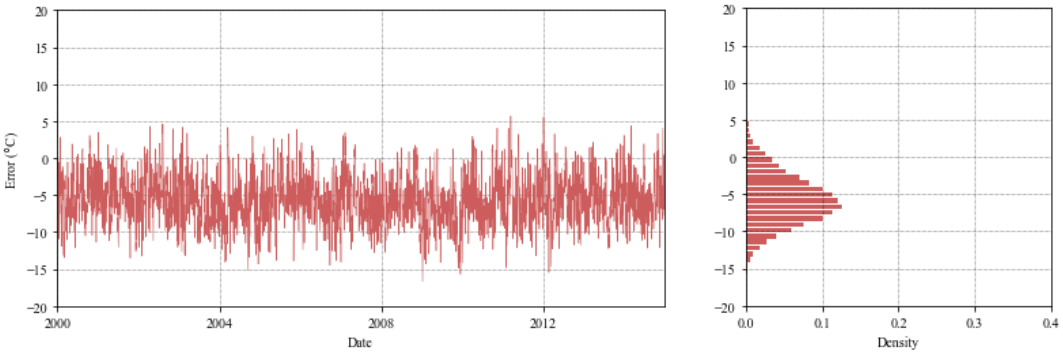
Fig. 9.1 Observations, predictions, and error between the two of average precipitation over the subdomain encompassing the UK for the years 2000-2014, with both time series data and associated histograms, displaying density, for those time series



(a) Time series and associated histogram of observations



(b) Time series and associated histogram of hindcasts



(c) Time series and associated histogram of the observation-hindcast error

Fig. 9.2 Observations, predictions, and error between the two of average temperature over the subdomain encompassing the UK for the years 2000-2014, with both time series data and associated histograms, displaying density, for those time series

its suitability for forcing a machine learning model, and hence that of GCM output data, and the confidence that can be attached to the output.

9.1.1 Quantile Mapping

The procedure for Quantile Mapping is to use the ratio between a selected quantile from the distribution of historical observations and historical model output, or hindcast, to correct all model output, both historical and future. Therefore, for some climatic variable, X, and a given GCM, the Quantile Mapping method, for the nth quantile, the ratio of the historical, $X_{Q_n:obs,h}$, and hindcast, $X_{Q_n:GCM,h}$, is expressed in Equation 9.1, resulting in the quantile corrected version of that variable, X_{QC} (Maraun and Widmann, 2018).

$$X_{QC} = X_{GCM} \cdot \frac{X_{Q_n:obs,h}}{X_{Q_n:GCM,h}} \tag{9.1}$$

Ordinarily, Quantile Mapping would be a global, linear correction; in other words it would be agnostic to location within the temporal domain and this would be equivalent to that correction being the raw data multiplied by some coefficient, c s.t. $x' = c \cdot x$. Instead, we will assume that the bias needs to be corrected seasonally, to better represent the seasonality of the climatic variables in question. Thus, the quantile mapping will be performed across the entire historical record by month (Piani et al., 2010).

With our aim being increasing the similarity between the observational data and the hindcast, a measure of closeness between those statistical distributions must first be defined. In addition to the Kullback–Leibler Divergence, as described in Chapter 3, the Wasserstein Distance will be included. Whilst the KLD is useful as a measure of the similarity between distributions in terms of the shape of that distribution, it doesn't reflect the distance between similar distributions, such as if two Gaussian distributions with the same standard deviation had different means; Wasserstein Distance, on the other hand, is a measure of the amount of 'energy' required to adjust a statistical distribution such that it is identical to another (Kolouri et al., 2017). The pth Wasserstein Distance, W_p, is expressed in terms of two distributions, U and V, where $U, V \in P_p(\mathbb{R}^d)$ with finite pth moment, in Equation 9.2. Both KLD and Wasserstein Distance lie within the interval $[0, \infty)$, where 0 indicates perfect agreement or similarity between the two distributions.

$$W_p(U,V) := \left(\underset{\pi_{X,Y} \in \Pi(U,V)}{inf} \mathbb{E}_\pi \left[||X-Y||^p \right] \right)^{\frac{1}{p}} \tag{9.2}$$

Where the set, $\Pi(U,V)$, of all couplings, $\pi_{X,Y}$, of U and V is defined by Equation 9.3:

$$\Pi(U,V) = \left\{ \pi_{X,Y} \in P(\mathbb{R}^d \times \mathbb{R}^d) \ : \ \pi_X = U \ \& \ \pi_Y = V \right\} \tag{9.3}$$

After Quantile Mapping is applied to the UK dataset, the KLD and Wasserstein Distance between the distributions of corrected versus uncorrected time series data and the observational data at each of the four test catchments are shown in Table 9.1. Also shown is a section of the corrected precipitation time series for the Severn catchment along with the change in error signal between the raw and bias corrected data, in Figure 9.3. The Quantile Mapping procedure, as one that is more a linear transform that acts to shift the distribution rather than alter its shape, results in a marked improvement in the Wasserstein Distance but a negligible improvement in the KLD. Whereas for temperature the KLD is already close to optimal, highlighting that temperature is well resolved within climate models, this is less the case for precipitation and, although Wasserstein Distance is improved, Quantile Mapping is perhaps less suitable for precipitation data in addressing the KLD discrepancy. Both the Wasserstein Distance and KLD are acceptable for the Quantile Mapped temperature and wind speed data.

Table 9.1 Wasserstein Distance and KLD for the original raw GCM data and the quantile mapped data at each of the four test catchments

Catchment	Data	Precipitation		Temperature		Windspeed	
		W	KLD	W	KLD	W	KLD
Severn at Haw Bridge	Raw Output	2.179	3.334	4.671	0.001	5.328	0.250
	Corrected	0.210	3.322	1.377	0.001	0.976	0.247
Thames at Kingston	Raw Output	1.967	4.627	5.319	0.001	3.882	0.257
	Corrected	0.098	4.608	1.648	0.001	1.112	0.255
Avon at Bathford	Raw Output	2.347	4.468	5.300	0.001	4.005	0.254
	Corrected	0.173	4.454	1.565	0.001	1.087	0.252
Exe at Pixton	Raw Output	2.429	4.120	5.052	0.001	4.302	0.244
	Corrected	0.204	4.107	1.756	0.001	1.161	0.243

9.1.2 Machine Correction

This section is more of a footnote as a machine learning framework to correct the bias and downscale the model output was investigated but the results were inconclusive; a Neural Process algorithm was applied to learn the mapping between the raw GCM data and observations but the KLD cost function exhibited no convergence and the R^2 value between the observations and hindcasts no improvement, remaining below 0 both before and after training.

The conclusion drawn here is that there is systematic error within the model that causes divergence from real observations and for which it is highly unlikely a transfer function exists, at least not within the same space. Whilst we do not rule out the possibility that one might be able to transform the datasets to another space and learn a transfer function within that space, it is perhaps not an avenue worth exploring given the nature of GCMs.

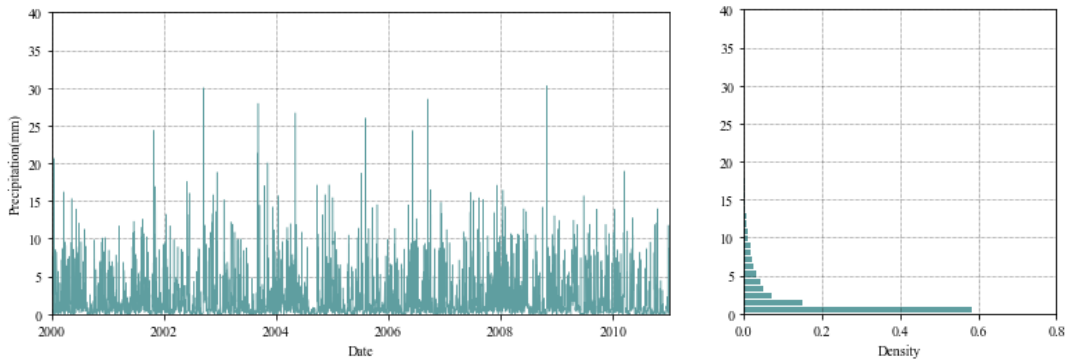
Ultimately, the bias corrected output of these methods is not identical to actual observations (Chen et al., 2012; Maraun, 2016) and nor should that be expected, given that climate state is not necessarily equivalent to weather. In the absence of Regional Climate Models forced by a GCM, for which the computational demand to drive an RCM for all subdivisions of the global domains would be excessive (McGregor, 1997), obtaining a dataset to force impact models that is representative without necessarily having perfect correlation with observations is the minimum for investigating the change in future trends and is what has been achieved to an extent.

9.2 Potential Scenarios

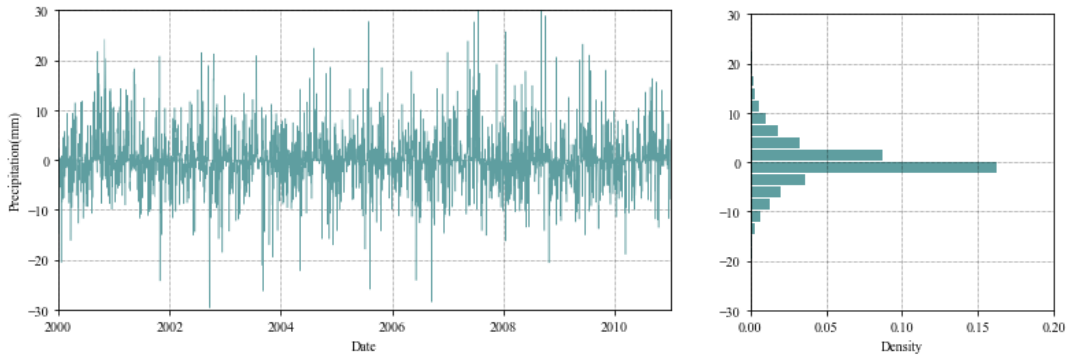
For each of the catchments, two separate approaches were adopted: the first being to train - or, in a sense, calibrate - the model using the historical observational data and then use the future climate data to force that model; the second is to repeat the former but use the MPI-ESM hindcasts as the training data. The observation-trained impact models will be referred to as IM-**O** and the hindcast-trained as IM-**H**. In both instances, the Neural Process model was fitted and the future projections used for the 4 selected SSPs.

Daily streamflow predictions generated with the yearly distribution, in terms of the mean and standard deviation, for all scenarios are shown for each of the four catchments using IM-**O** in Figure 9.4 and using IM-**H** in Figure 9.5.

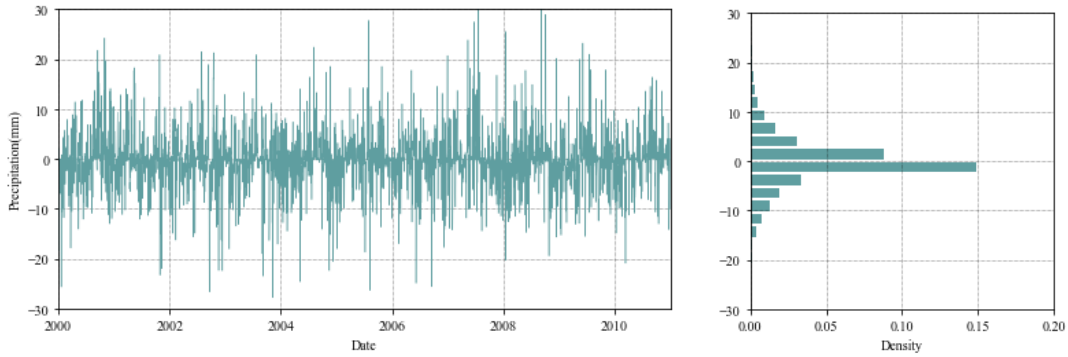
The first observation to be made is that the projections made by models with different training frameworks are significantly different for the same input; the growth in streamflow is more modest across IM-**O** projections, with the exception of the Thames at Kingston, when compared with IM-**H** projections and the mean signal is also more volatile for the latter. If



(a) Time series and associated histogram of observations



(b) Time series and associated histogram of the observation-hindcast error



(c) Time series and associated histogram of the bias corrected observation-hindcast error

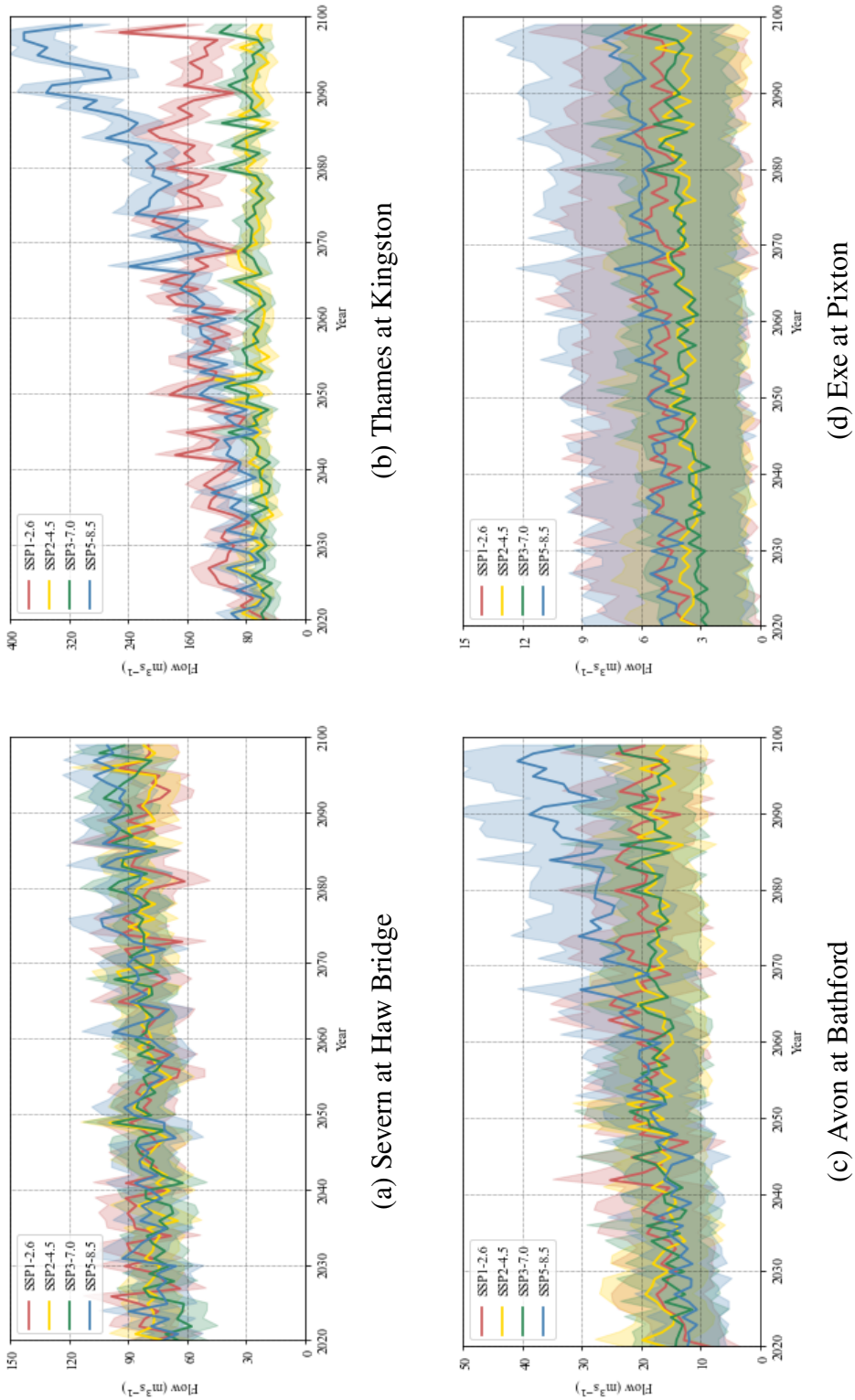
Fig. 9.3 Observations and the error signal between the observations and uncorrected hindcasts, middle, and between the observations and bias corrected hindcasts averaged over the River Severn at Haw Bridge catchment for the years 2000-2014, with both time series data and associated histograms, displaying density, for those time series

one were to choose which projections were more likely, then deliberating as to which set of projections is more accurate is best left unanswered for now and replaced with a more apt question: that of which is the more inherently flawed approach, and to which there is a more apparent answer. Given that there is still significant bias within Quantile Mapped input data, then the IM-**H** model is learning unrepresentative dependencies; this was clear at the outset but the approach was nevertheless pursued to analyse the change in model response to the change in the input, partly to further underscore the fallacy of this approach but also to investigate the nature of that response.

Recall from Chapter 1 that the global precipitation increase in intensity was expected to be up to 50% or more, which included a similar increase for the UK, as in Figure 1.4. So if more precipitation is expected then streamflow should be expected to increase, as we consider this the primary driver of streamflow as was highlighted in Chapter 4. Under the higher warming scenarios, the output of the streamflow model exhibits greater variability with a greater mean flow and, conversely, there is less change in the streamflow distribution for the SSP1-2.6 scenario. Due to climate being expected to remain closer to present day for the SSP1-2.6 scenario and diverging the most for the SSP5-8.5 scenario, the high level trend in streamflow distributions from the model is certainly believable on the surface but temptation must be ignored.

Improbable instances, with regards to river capacity, appear in the future flow scenarios that cast a pall over the direct credibility of predictions. For example, in the historical period 1979 to 2019 for the River Severn, the highest flow event recorded was approximately $1250\text{m}^3\text{s}^{-1}$, whereas under the SSP5-8.5 scenario scenario the highest flow event output by the model is approximately $6000\text{m}^3\text{s}^{-1}$; for the 2007 flow event, there was significant inundation at Tewkesbury (Coates et al., 2020), approximately 7.5km upstream from the gauge site, so the capacity of the river is most likely unable too support a streamflow of $6000\text{m}^3\text{s}^{-1}$. The implication being that this would necessarily be accompanied by significant flooding with most of this flow exceeding the maximum capacity of the river and, thus, such events need to be separated into the amount of precipitation and subsequent runoff within the catchment and the actual corresponding streamflow. Above a certain threshold, estimated flows are likely in excess of the river's capacity and therefore not realistic.

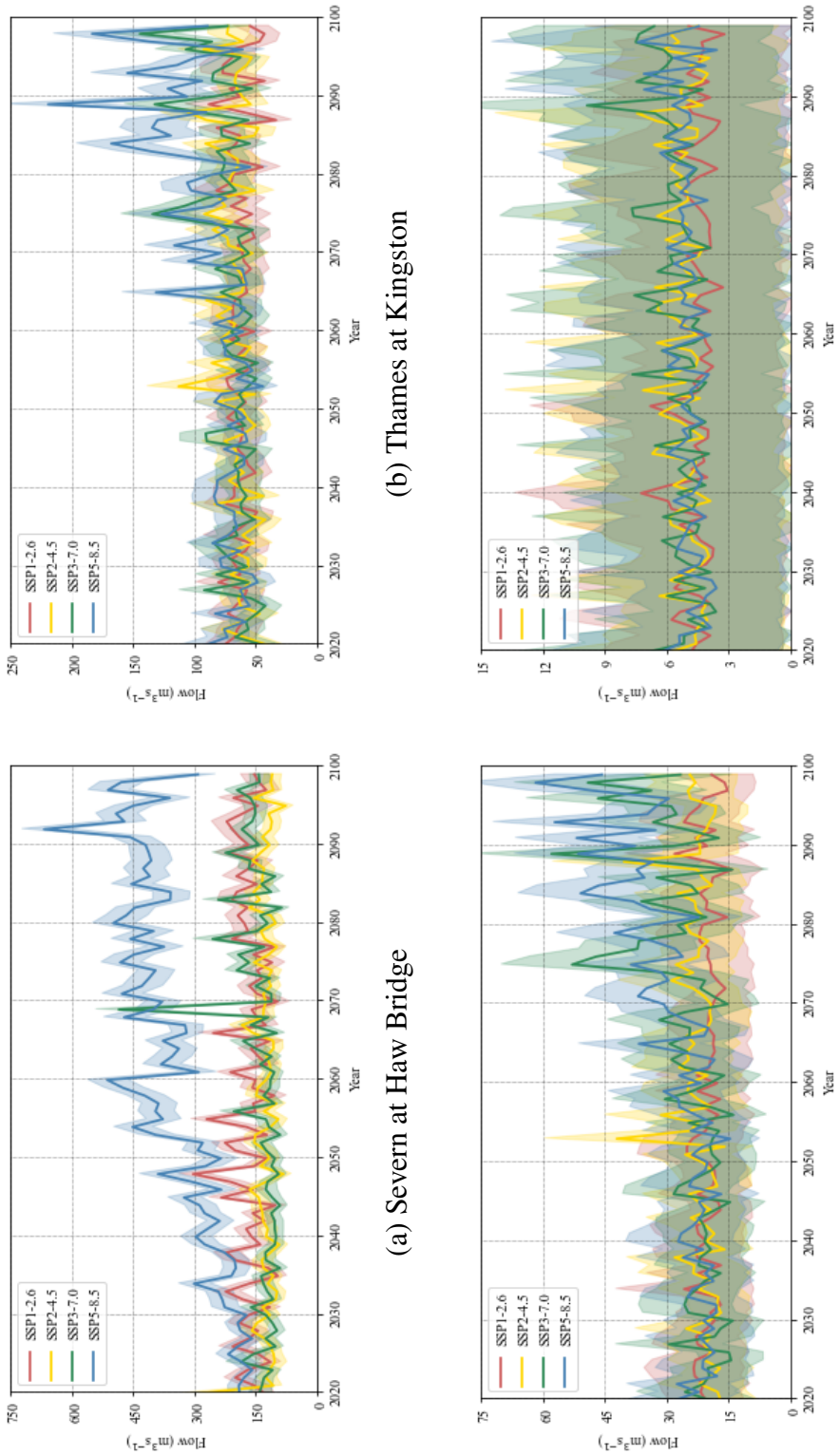
And so, if the output from IM-**H** is rejected, then the fallback position would be to assume that the output IM-**O** is more accurate and that it might provide robust insight into future streamflow distribution. Ergo, we much weigh its output against expectation, especially with regards to the shape of the climate in the future and how we expect the world to be.



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 9.4 Projected distribution of streamflow by year using IM-**O** under four SSP scenarios with the mean for each scenario represented by the coloured, solid lines and the range between two standard deviations of the respective mean by the same coloured, filled area

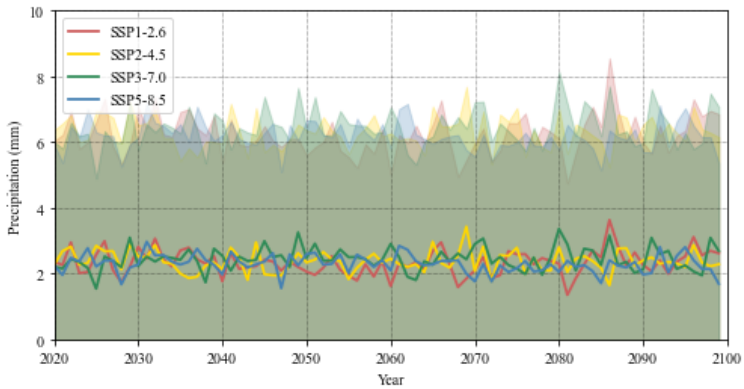


(a) Severn at Haw Bridge (b) Thames at Kingston

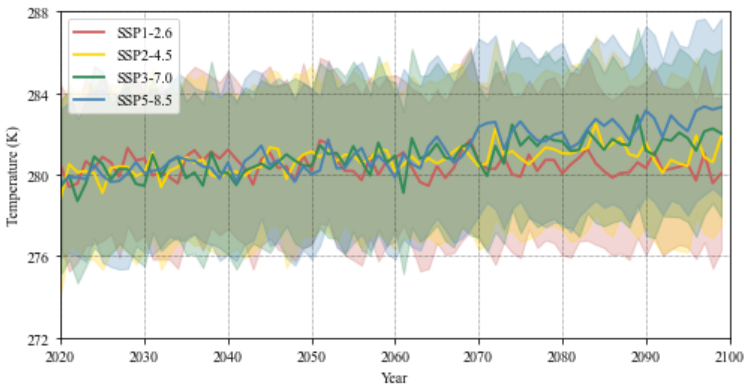
(c) Avon at Bathford (d) Exe at Pixton

Fig. 9.5 Projected distribution of streamflow by year using IM-**H** under four SSP scenarios with the mean for each scenario represented by the coloured, solid lines and the range between two standard deviations of the respective mean by the same coloured, filled area

For two of the catchments, there is negligible difference between the different scenarios, with all resulting in a similar shift in the streamflow distribution over time; however, for the Thames and Avon the SSP5-8.5 scenario results in a markedly higher mean with greater variance by 2100 with the next highest flow distribution occurring under the SSP1-2.6 scenario. These results might appear inconsistent with our prior belief that the greater warming scenarios for the United Kingdom would result in more variable streamflow and a higher average, thus warranting a closer inspection of the MPI-ESM projections of precipitation and temperature. In Figure 9.6, distributions of precipitation and temperature, again in terms of mean and variance, by year are shown for time period of the study.



(a) Precipitation



(b) Temperature

Fig. 9.6 Projected distribution of selected climatic variables by year under four SSP scenarios with the mean for each scenario represented by the coloured, solid lines and the range between two standard deviations of the respective mean by the transparent filled area of the same colour

Whilst there is a general increasing trend for temperature, in particular for the higher emissions scenarios, as is in line with expectations, the same is not the case for precipitation. In fact, the same negligible difference in domain average precipitation mirrors that of the average streamflow. The MPI-ESM data does not suggest a change in precipitative trends and the transformation applied to the GCM data was insensitive to temporal trends, meaning that it could not have eliminated long-term temporal fluctuations. Therefore, if one were to use the MPI-ESM output as the sole forcing for modelling potential impact, any infrastructure or policy decisions based on said assessment would account for limited change, given that the response of IM-**O** to its forcing appears valid. If we were to assume that this is not the case, then the alternative would be to use additional data for the purposes of comparison or taking a more holistic approach.

9.2.1 An Alternative Forcing

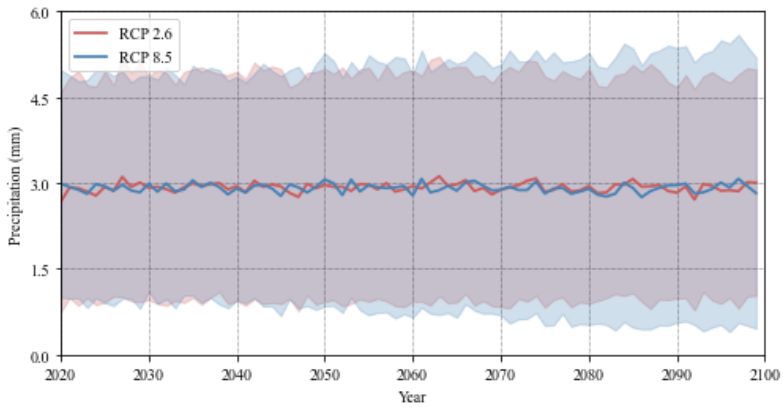
As described in Chapter 3, the Met Office UK Climate Projections 2018 (UKCP18) could be a forcing input to the streamflow model that offers a comparison point in spite of its derivation from CMIP5 due to the additional postprocessing done to attain high-resolution spatially-coherent future climate projections (Met Office Hadley Centre, 2019). With it, we may be able to either corroborate or highlight the discrepancy between the output of IM-**O** forced by MPI-ESM, given the skew of UKCP18 towards modelling accuracy over the UK region.

The same approach is adopted as before, using Quantile Mapping with similar relative improvement in the Wasserstein Distance prior to forcing the pretrained IM-**O** model. The projected yearly distributions of precipitation and temperature are presented in Figure 9.7 and the resulting yearly distributions for the four test catchments under the two available RCP scenarios, being those with 2.6 Wm^{-2} and 8.5 Wm^{-2} radiative forcing, are shown in Figure 9.8.

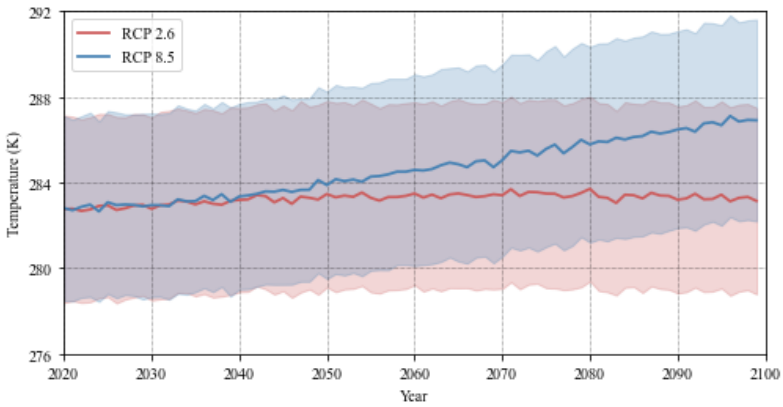
Results from the higher radiative forcing scenario, RCP 8.5, exhibit greater variability and a marginally higher mean for streamflow distributions using UKCP18, with the exception of the Thames, where the different scenarios both would suggest an increasing trend in mean streamflow by approximately 60%. Again, considering the input to the model, in terms of the relatively greater increase in mean temperature and similarity in mean precipitation between the RCP scenarios, the minor impact on divergence between streamflow distributions necessarily follows, given model sensitivity to precipitation.

The picture provided by using the UKCP18 data is similar to that when using MPI-ESM data, where the change in precipitation is more modest than for temperature and the resulting response from IM-**O** between scenarios is likewise perhaps modest, with the difference in

streamflow distribution between those scenarios for all four catchments being limited by 2100 albeit to a different extent. Though the pattern in impact may be congruous, the suitability of a climate projection being used to provide an assessment of streamflow, being sensitive to localised weather phenomena, is conversely incongruous.

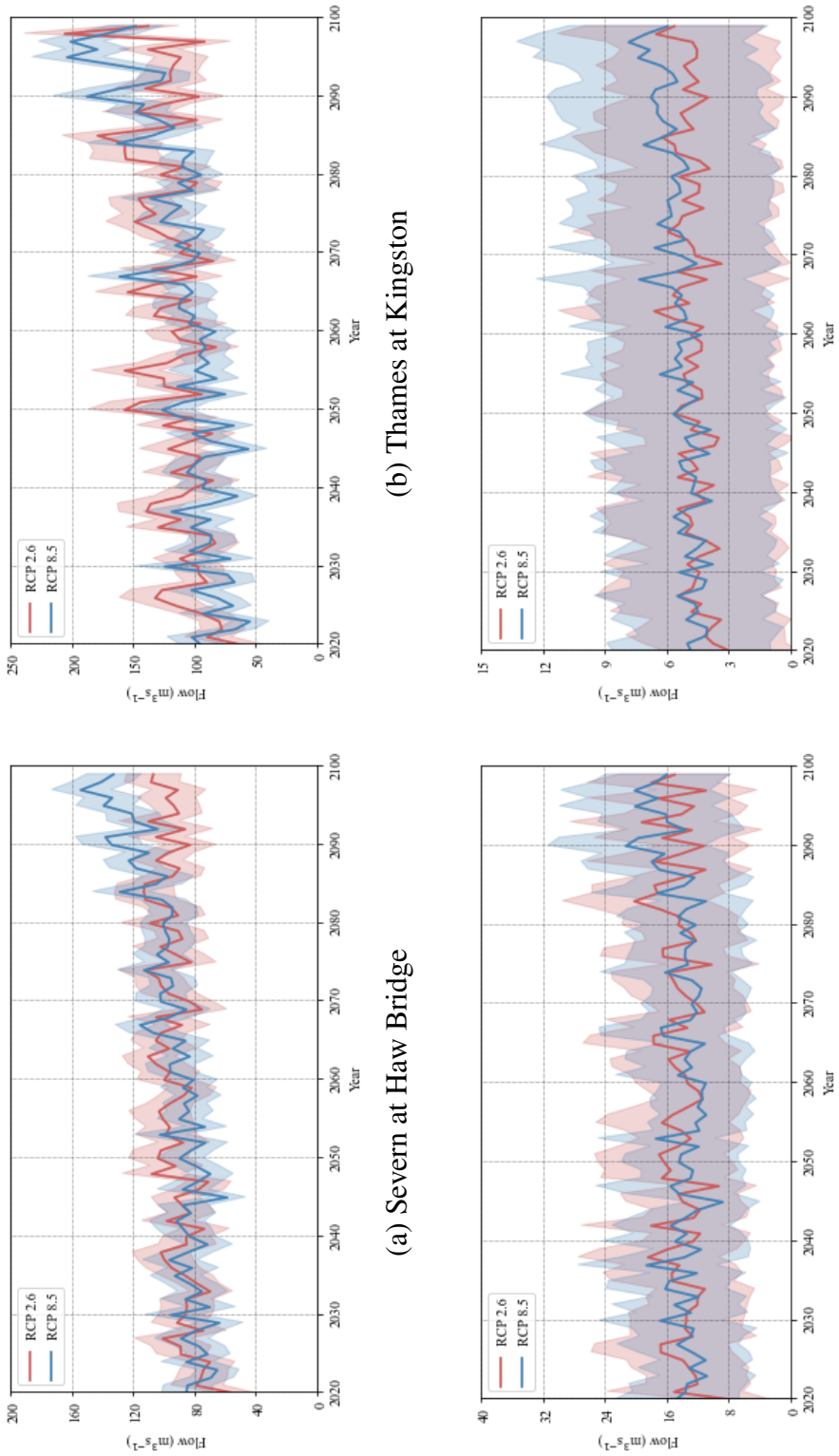


(a) Precipitation



(b) Temperature

Fig. 9.7 Projected distribution of selected climatic variables by year under two RCP scenarios with the mean for each scenario represented by the coloured, solid lines and the range between two standard deviations of the respective mean by the transparent filled area of the same colour



(a) Severn at Haw Bridge (b) Thames at Kingston

(c) Avon at Bathford (d) Exe at Pixton

Fig. 9.8 Projected distribution of streamflow by year using IM-**H** under two RCP scenarios with the mean for each scenario represented by the coloured, solid lines and the range between two standard deviations of the respective mean by the same coloured, filled area

9.2.2 Predictive Utility

Reflecting on Quantile Mapping as a means to alter data for use in a downstream model, we can only conclude that this is a sub-optimal procedure when used on its own. Although the correction procedure brought the data more in line with the statistical distributions, the reduction in the discrepancy between historical observations and hindcasts was modest.

Of the two initial approaches, one, IM-**H**, is undoubtedly incapable of representing the correct pattern mapping, whilst the other, IM-**O**, is learning a suitable representation but the results may not necessarily be reflective of outcomes when we consider that the change in climate does not simulate at a more granular level the subsequent change in weather phenomena. Thus, there is an important point to be made with regards to the desired output of a model not being prioritised over the quality of the model itself. In this instance, the output from IM-**H** ought to be dismissed, whilst that of IM-**O** requires bolstering. As multiple models are used as part of CMIP6, so too should multiple models be used; having said that, the strategy adopted for combining multiple models into an ensemble requires careful consideration, given the difference in how processes are represented (Tebaldi and Knutti, 2007).

One must also keep in mind the purpose of climate modelling and that, in keeping with the principle that any predictions generated are not equivalent to future behaviour but representative of hypothetical scenarios, the picture that they paint should not be interpreted literally, particularly when it comes to subgrid scale weather phenomena. Ergo, the direct forcing of a machine learning model using the output of a GCM, modified or not, is not something we recommend here.

9.3 Future Direction

An alternative proposition is required, one that shows promise in surpassing the caveats attached here to using bias corrected GCM data to force impact models, given that the potential impacts of climate change are very much in need of quantification. We again underscore the difference between climate and weather and the need for a paradigm focused on weather patterns under climate scenarios but believe the notion of the comparison between statistical distributions to be useful to an extent.

If we let the current climate be a seasonally varying statistical model, $f(\Theta_o|t)$, with parameters, Θ_o fitted to the respective seasonal weather and the simulated climate be a similar statistical model, $g(\Theta_G|t)$, with parameters, Θ_G, then we can learn a function mapping, χ, from the simulated climate model to the observed climate using hindcasts, as per Equation 9.4.

$$\chi : g(\Theta_G|t) \mapsto f(\Theta_o|t) \tag{9.4}$$

Assuming that the mapping is valid, it can be used to map the forecasts to future climate, yielding a climate distribution for future climate under the different emissions scenarios. A parameterised statistical distribution could enable the use of physics informed time series models (Kashinath et al., 2021; Lim and Zohren, 2021), where the changing parameters for the climate statistical distribution could then serve as the input features with which to force a generative weather model, producing weather variable time series of arbitrary length.

An important aspect of this framework is that a realistic representation of weather in this case would need to account for features such as the balance and variance between wet and dry periods, which is often underrepresented through traditional approaches (Chen et al., 2012). There is no shortage of data with which to train the generative model, given the large historical meteorological datasets such as ERA5 used throughout this thesis, so one would anticipate that the required representation is obtained.

Thus, the learned mapping, χ, can force the parameters, Θ_o, that define the projected future climate distribution which can then be used to generate future weather time series of a given length. The relative efficiency of computation is such that multiples of such weather simulations can be created with these weather simulations then being used in turn to force the hydrological machine learning model. We then envisage being able to take the evaluated output of the hydrological impact model to create a further distribution of streamflow, given the high number of weather simulations that would be generated, providing a richer, probabilistic view of the likely seasonal change with respect to time.

9.4 Concluding Remarks

Our primary finding here is that linear transforms of General Circulation Model data are inappropriate for creating projections of future impact under different emissions scenarios, despite their ubiquity. Our reasoning is due to the fact that meteorological observations and climate simulations are inherently different and treating the difference between the two as an error term that can be easily corrected does not properly treat the underlying physics in either situation.

Instead, we have proposed a new hypothesis for learning a mapping between observations and climate model data to be used to force a generative time series model. Complementary to the prior point is that machine learning models are likely suitable for creating impact studies,

provided that the training data is scrutinised and is of similarly high standard to that used to force them.

Chapter 10

Summary & Conclusions

10.1 Chapter Summarisation

10.1.1 Chapter 4 - Establishing a Basis

The first output from this thesis was a streamflow modelling approach constructed from first principles and one that was receptive to climatic variables only, ones that could be extracted from global meteorological and climate modelling products. Through that framework, we demonstrated the need for nonlinear modelling capability and, by virtue of machine learning parameter optimisation, also that the system response is most sensitive to precipitation and lower temperatures than other variables.

By leveraging the innate properties of machine learning algorithms, specifically their being black box by nature and being able to internalise interim processes, we showed that variables describing system state can be subsumed into the model. This enabled us to define a compressed feature set with statistical proxy variables for soil moisture, alleviating the need for those measurements internal to a catchment.

10.1.2 Chapter 5 - A Model Comparison

With an appropriate feature set, the next step was to examine the effect of machine learning architecture on the model's performance. The complexity of the Multi-Layer Perceptron was steadily increased and whilst the expectation might have been that superior performance would ensue with every refinement there are caveats with increasing the complexity. For example, as a time series problem, sequence models like the Temporal Convolutional Neural Network would appear more naturally suited but the base MLP was model architectures are easier to fit and train. We also note that if an understanding between inputs and outputs can

be expressed, then that a priori knowledge about a system and how it behaves opens up the use of powerful and elegant Gaussian Process models. The hybrid Neural Process models offered superlative performance, whilst being balanced between ease of implementation, computational complexity, and able to learn rapidly from data.

A related outcome is that the choice of model architecture is very much secondary to the feature engineering done in the preceding chapter. Our results for basic machine learning models was comparable to that of the more advanced techniques. Essentially, domain knowledge remains a key component in machine learning and the quality or choice of data is paramount. This was further evidenced by the fact that no model was able to overcome the lack of information in driving the River Exe Summer and Winter peaks of 2012.

Extremes are, arguably, of more interest than general hydrological behaviour; storms, flooding, and drought are events with high socioeconomic and environmental impact. Ergo, we sought to and did develop a metric that was more relevant to the quantification of performance around these extremes. This metric, being a non parametric adaptation of R^2 and that in turn being a normalisation of the Mean Squared Error, could have further potential as a cost function for optimising machine learning models, driving performance higher towards the modelling of extreme events.

10.1.3 Chapter 6 - Generalising Across Catchments

The framework for single catchments, developed in terms of the feature engineering and model architecture, we then extended towards multi-catchment modelling. Our aim was to have a model capable of learning from specific geographies and then extrapolate with a high level of accuracy towards others, based on easily accessible catchment descriptors. Whilst this framework was only tested across UK geographies, owing to our data ingress strategy, it has demonstrable potential for global expansion through the inclusion of additional permutations, in particular the topographical and meteorological.

A critical aspect of the model here is the expression of beliefs about human influences on a system through a simplified approach that yields comparatively high levels of performance for heavily affected, yet understudied, systems. Our exemplar here is the River Darent at Hawley; the conventional approach without accounting for human influence generates unusable results. As part of an ensemble approach, the majority of human influences and, therefore, a far wider range of catchments can be modelled accurately.

10.1.4 Chapter 7 - Sequential Process Models

In prior chapters, we commented on the high performance of the Neural Process; combining this with the notion that the structure of a neural network is a means of encoding a belief about the behaviour of a modelled system, almost analogous to a Gaussian Process kernel, prompted our investigation of Sequential Processes. Replacing the encoder and decoder networks within the Neural Process with Recurrent Neural Networks provided an improvement in performance over both the standalone RNN model and over the base Neural Processes for this application. With a Temporal Convolutional Neural Network, however, we were unable to realise any improvement.

Our conclusion here is that the Neural Process architecture may accommodate a wide range of encoder and decoder networks that are better suited to the task at hand. However, as was the case with the modelling approaches outlined in Chapter 5, a balance between model complexity and tailoring the approach needs to be struck.

10.1.5 Chapter 8 - Storm Prediction

A machine learning algorithm, utilising architectures developed for computer vision tasks, was applied to gridded circulation data in order to identify patterns corresponding to storm activity. We made further refinements to the model, enabling predictions for precipitative impact of reasonable accuracy whilst efforts to localise storm epicentres within a domain resulted in middling success, to be further improved upon with more appropriate data labelling.

The initial aim of being able to project the forward movement and impact of storms was not met but we have provided a route towards achieving that aim with a incremental improvement upon the work completed here. Additionally, using the predicted precipitative impact from these extreme weather phenomena to force a hydrological model and obtain a response is an elementary step to be undertaken for storm and catchment modelling with overlapping spatial data.

10.1.6 Chapter 9 - Impact Modelling

Linear transforms were explored but our belief is that they are unsuitable for simulating the impacts of climate change under varying emissions scenarios. We, instead, proposed another route that is more in keeping with the content of this thesis and one that does not run counter to the underlying physics of General Circulation Models.

As with previous chapters, the quality of data going into a model is critical and scrutinising the outputs, especially with regards to their alignment with expectations, is mandatory. On the inputs, those used to force a trained model ought to be comparable with those used in that training process and our work here demonstrates the need for comparative analysis that is fit for context; for example, a linear transform may improve a metric such as Wassertein Distance but does not address Kullback–Leibler Divergence.

10.2 Holistic Conclusion

At the beginning of this thesis, we highlighted the need for superior predictive storm and river behaviour modelling in the face of anthropogenic climate change; this had to be balanced against the fact that global data coverage is inequitable. This latter issue prevents the deployment of models requiring situational calibration and is made more acute when considering the uneven impact of varying precipitation and temperature trends under different emissions scenarios.

None of these hydrological elements or their impacts exist in isolation, though; there is a clear process flow through them and the overarching theme through which they are woven is where the true value of this thesis lies. By concerning ourselves with extremes and being able to make predictions thereof with minimal data, either by considered selection of variables and proxies or through the transfer of learned patterns from well monitored regions, we have arrived at a framework for hydrological climate impact that is both flexible and robust.

Foremost, and more formally, this thesis presents a route towards generating hydrological impact studies under varying climate scenarios, including being able to identify extreme weather phenomena and predict their forcing for hydrological extremes, with a global remit, unhindered by regional data sparsity.

10.3 Final Remarks

When measured against the stated aims of the research, as described in Chapter 1, the majority of its components are complete; for those that either need further improvement or investigation, we have outlined a route that would lead towards the full attainment of all research objectives. The expansive scope of the work was both an aspect of strength and a hindrance but, ultimately, we feel the overall objective has been addressed.

Across this entire body of work, one of the most important challenges has been that of expectation; expecting state of the art performance necessarily requires superior quality and availability of data. A caveat must also be placed upon any future predictions, in terms of

what is realised through modelling and what is expected. If the onus is placed on obtaining results that match expectations, then the reality might be occluded.

As for whether or not machine learning is appropriate for use in the context of environmental data science, specifically for investigating hydrological phenomena, our belief is that it is, dependent on its suitable and justifiable application guided by domain knowledge. State of the art performance can be achieved in most tasks, such as creating a generalising model capable of learning human interactions in the absence of data, and for those where it was not, such as storm localisation, the refinements outlined in the relevant sections ought to move them into that same performance bracket.

Finally, as the challenges of climate change become more evident and the time gap between delivering actionable information and the action itself to counter the effects of climate change shortens, machine learning approaches, being computationally efficient and able to generalise in the face of data paucity, can be leveraged. Thus, that required actionable information can be delivered, enabling prompt policy and infrastructure responses.

References

Abbott, M., Bathurst, J., Cunge, J., O'Connell, P., and Rasmussen, J. (1986a). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1-2):45–59.

Abbott, M., Bathurst, J., Cunge, J., O'Connell, P., and Rasmussen, J. (1986b). An introduction to the European Hydrological System — Systeme Hydrologique Europeen, "SHE", 2: Structure of a physically-based, distributed modelling system. *Journal of Hydrology*, 87(1-2):61–77.

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., and Wilby, R. L. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography: Earth and Environment*, 36(4):480–513.

Ackerman, F., DeCanio, S. J., Howarth, R. B., and Sheeran, K. (2009). Limitations of integrated assessment models of climate change. *Climatic Change*, 95(3-4):297–315.

Aichouri, I., Hani, A., Bougherira, N., Djabri, L., Chaffai, H., and Lallahem, S. (2015). River Flow Model Using Artificial Neural Networks. *Energy Procedia*, 74:1007–1014.

Ali, S. and Shahbaz, M. (2020). Streamflow forecasting by modeling the rainfall–streamflow relationship using artificial neural networks. *Modeling Earth Systems and Environment*, 6(3):1645–1656.

Allard, R. F. (2021). Climate Change Adaptation: Infrastructure and Extreme Weather. In Leal Filho, W., Azul, A. M., Brandli, L., Lange Salvia, A., and Wall, T., editors, *Industry, Innovation and Infrastructure*, pages 105–116. Springer International Publishing, Cham. Series Title: Encyclopedia of the UN Sustainable Development Goals.

Allen, J. T. (2018). Climate Change and Severe Thunderstorms. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press.

Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D. C., Wilkinson, J., Phillips, T., Byrne, J., Tietsche, S., Sarojini, B. B., Blanchard-Wrigglesworth, E., Aksenov, Y., Downie, R., and Shuckburgh, E. (2021). Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(1):5124.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R. (1998). LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART I: MODEL DEVELOPMENT. *Journal of the American Water Resources Association*, 34(1):73–89.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000). Artificial Neural Networks in Hydrology. I: Preliminary Concepts. *Journal of Hydrologic Engineering*, 5(2):115–123.

Auld, H. E. (2008). Adaptation by design: the impact of changing climate on infrastructure. *Journal of Public Works & Infrastructure*, 1(3):276–288.

Ayzel, G. and Heistermann, M. (2021). The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. *Computers & Geosciences*, 149:104708.

Balog, I., Caputo, G., Iatauro, D., Signoretti, P., and Spinelli, F. (2023). Downscaling of Hourly Climate Data for the Assessment of Building Energy Performance. *Sustainability*, 15(3):2762.

Bartens, A. and Haberlandt, U. (2021). Flood frequency analysis using mean daily flows vs. instantaneous peak flows. preprint, Catchment hydrology/Stochastic approaches.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Besaw, L. E., Rizzo, D. M., Bierman, P. R., and Hackett, W. R. (2010). Advances in ungauged streamflow prediction using artificial neural networks. *Journal of Hydrology*, 386(1-4):27–37.

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, 5(1):1–12.

Bian, G., Zhang, J., Chen, J., Song, M., He, R., Liu, C., Liu, Y., Bao, Z., Lin, Q., and Wang, G. (2021). Projecting Hydrological Responses to Climate Change Using CMIP6 Climate Scenarios for the Upper Huai River Basin, China. *Frontiers in Environmental Science*, 9:759–547.

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, New York, NY, softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009) edition.

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., Aronica, G. T., Bilibashi, A., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Fiala, K., Frolova, N., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Rogger, M., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Viglione, A., Volpi, E., Wilson, D., Zaimi, K., and Živković, N. (2017). Changing climate shifts timing of European floods. *Science*, 357(6351):588–590.

Boer, G. J. (1993). Climate change and the regulation of the surface moisture and energy budgets. *Climate Dynamics*, 8(5):225–239.

Bosler, P. A., Roesler, E. L., Taylor, M. A., and Mundt, M. R. (2016). Stride Search: a general algorithm for storm detection in high-resolution climate data. *Geoscientific Model Development*, 9(4):1383–1398.

British Geological Survey (2007). BGS Geology 625K. Database - https://www.bgs.ac.uk/datasets/bgs-geology-625k-digmapgb/.

Brochu, E., Brochu, T., and Freitas, N. D. (2010). A Bayesian Interactive Optimization Approach to Procedural Animation Design. *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*, page 10 pages. Artwork Size: 10 pages ISBN: 9783905674279 Publisher: The Eurographics Association.

Brown, C. (2010). The End of Reliability. *Journal of Water Resources Planning and Management*, 136(2):143–145.

Brunner, P. and Simmons, C. T. (2012). HydroGeoSphere: A Fully Integrated, Physically Based Hydrological Model. *Ground Water*, 50(2):170–176.

Buchhorn, M., Smets, B., Bertels, L., Roo, B. D., Lesiv, M., Tsendbazar, N.-E., Herold, M., and Fritz, S. (2020). Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe.

Butchart, S. H. M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J. P. W., Almond, R. E. A., Baillie, J. E. M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K. E., Carr, G. M., Chanson, J., Chenery, A. M., Csirke, J., Davidson, N. C., Dentener, F., Foster, M., Galli, A., Galloway, J. N., Genovesi, P., Gregory, R. D., Hockings, M., Kapos, V., Lamarque, J.-F., Leverington, F., Loh, J., McGeoch, M. A., McRae, L., Minasyan, A., Morcillo, M. H., Oldfield, T. E. E., Pauly, D., Quader, S., Revenga, C., Sauer, J. R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S. N., Symes, A., Tierney, M., Tyrrell, T. D., Vié, J.-C., and Watson, R. (2010). Global Biodiversity: Indicators of Recent Declines. *Science*, 328(5982):1164–1168.

Buytaert, W., Célleri, R., and Timbe, L. (2009). Predicting climate change impacts on water resources in the tropical Andes: Effects of GCM uncertainty: CLIMATE CHANGE IMPACTS ON WATER RESOURCES. *Geophysical Research Letters*, 36(7):n/a–n/a.

Calver, A., Stewart, E., and Goodsell, G. (2009). Comparative analysis of statistical and catchment modelling approaches to river flood frequency estimation: River flood frequency estimation. *Journal of Flood Risk Management*, 2(1):24–31.

Camargo, S. J. (2013). Global and Regional Aspects of Tropical Cyclone Activity in the CMIP5 Models. *Journal of Climate*, 26(24):9880–9902.

Cangialosi, J. P., Blake, E., DeMaria, M., Penny, A., Latto, A., Rappaport, E., and Tallapragada, V. (2020). Recent Progress in Tropical Cyclone Intensity Forecasting at the National Hurricane Center. *Weather and Forecasting*, 35(5):1913–1922.

Canuti, P. and Moisello, U. (1982). Relationship between the yearly maxima of peak and daily discharge for some basins in Tuscany. *Hydrological Sciences Journal*, 27(2):111–128.

Cavicchia, L., Ascenso, G., Scoccimarro, E., Castelletti, A., Giuliani, M., and Gualdi, S. (2022). Tropical cyclone genesis potential in CMIP6 climate models. other, display.

Chadwick, A. J., Morfett, J., and Borthwick, M. (1998). *Hydraulics in civil and environmental engineering*. E & FN Spon, London ; New York, 3rd edition.

Challinor, A. J., Osborne, T., Morse, A., Shaffrey, L., Wheeler, T., Weller, H., and Vidale, P. L. (2009). Methods and Resources for Climate Impacts Research: Achieving Synergy. *Bulletin of the American Meteorological Society*, 90(6):836–848.

Chan, J. C. L. (2000). Tropical Cyclone Activity over the Western North Pacific Associated with El Niño and La Niña Events. *Journal of Climate*, 13(16):2960 – 2972.

Changnon, S. A., Pielke, R. A., Changnon, D., Sylves, R. T., and Pulwarty, R. (2000). Human Factors Explain the Increased Losses from Weather and Climate Extremes. *Bulletin of the American Meteorological Society*, 81(3):437–442.

Chen, G. and Tam, C.-Y. (2010). Different impacts of two kinds of Pacific Ocean warming on tropical cyclone frequency over the western North Pacific: Impacts of ocean warming on TC frequency. *Geophysical Research Letters*, 37(1).

Chen, J., Brissette, F. P., Chaumont, D., and Braun, M. (2013). Performance and uncertainty evaluation of empirical downscaling methods in quantifying the climate change impacts on hydrology over two North American river basins. *Journal of Hydrology*, 479:200–214.

Chen, J., Brissette, F. P., and Leconte, R. (2012). Coupling statistical and dynamical methods for spatial downscaling of precipitation. *Climatic Change*, 114(3-4):509–526.

Chen, J., Lin, S., Magnusson, L., Bender, M., Chen, X., Zhou, L., Xiang, B., Rees, S., Morin, M., and Harris, L. (2019). Advancements in Hurricane Prediction With NOAA's Next-Generation Forecast System. *Geophysical Research Letters*, 46(8):4495–4501.

Chen, S., Ren, Y., Friedrich, D., Yu, Z., and Yu, J. (2020). Sensitivity analysis to reduce duplicated features in ANN training for district heat demand prediction. *Energy and AI*, 2:100028.

Chia, H. H. and Ropelewski, C. F. (2002). The Interannual Variability in the Genesis Location of Tropical Cyclones in the Northwest Pacific. *Journal of Climate*, 15(20):2934 – 2944.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Church, J., Clark, P., Cazenave, A., Gregory, J., Jevrejeva, S., Levermann, A., Merrifield, M., Milne, G., Nerem, R., Nunn, P., Payne, A., Pfeffer, W., Stammer, D., and Unnikrishnan, A. (2013). Sea Level Change. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Coates, G., Alharbi, M., Li, C., Ahilan, S., and Wright, N. (2020). Evaluating the operational resilience of small and medium-sized enterprises to flooding using a computational modelling and simulation approach: a case study of the 2007 flood in Tewkesbury. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2168):20190210.

Cornelissen, T., Diekkrüger, B., and Bogena, H. (2016). Using High-Resolution Data to Test Parameter Sensitivity of the Distributed Hydrological Model HydroGeoSphere. *Water*, 8(5):202.

Cornwall, W. (2021). Europe's deadly floods leave scientists stunned. *Science*, 373(6553):372–373.

D. G. Morris and R. W. Flavin (1990). A digital terrain model for hydrology. *Proc 4th International Symposium on Spatial Data Handling*, 1:250–262.

D. G. Morris and R. W. Flavin (1994). Sub-set of UK 50 m by 50 m hydrological digital terrain model grids. Technical report, NERC, Institute of Hydrology, Wallingford.

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, 50(3):885–900.

Dawson, C., Abrahart, R., Shamseldin, A., and Wilby, R. (2006). Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology*, 319(1-4):391–409.

De Boor, C. (2001). *A Practical Guide to Splines*. Number v. 27 in Applied mathematical sciences. Springer, New York, rev. ed edition.

Devia, G. K., Ganasri, B., and Dwarakish, G. (2015). A Review on Hydrological Models. *Aquatic Procedia*, 4:1001–1007.

Ding, J., Haberlandt, U., and Dietrich, J. (2015). Estimation of the instantaneous peak flow from maximum daily flow: a comparison of three methods. *Hydrology Research*, 46(5):671–688.

Doocy, S., Daniels, A., Murray, S., and Kirsch, T. D. (2013). The Human Impact of Floods: a Historical Review of Events 1980-2009 and Systematic Literature Review. *PLoS Currents*.

Dooge, J. C. I. (1973). *Linear Theory of Hydrologic Systems*. Technical bulletin (United States. Dept. of Agriculture) ; no. 1468. Agricultural Research Service, U.S. Department of Agriculture.

Dottori, F., Szewczyk, W., Ciscar, J.-C., Zhao, F., Alfieri, L., Hirabayashi, Y., Bianchi, A., Mongelli, I., Frieler, K., Betts, R. A., and Feyen, L. (2018). Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change*, 8(9):781–786.

Dotzek, N., Groenemeijer, P., Feuerstein, B., and Holzer, A. M. (2009). Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. *Atmospheric Research*, 93(1-3):575–586. Publisher: Elsevier.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. (2011). Additive Gaussian Processes. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Edwards, R., Allen, J. T., and Carbin, G. W. (2018). Reliability and Climatological Impacts of Convective Wind Estimations. *Journal of Applied Meteorology and Climatology*, 57(8):1825–1845.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J. (2012). HESS Opinions "Should we apply bias correction to global and regional climate model data?". *Hydrology and Earth System Sciences*, 16(9):3391–3404.

Emanuel, K. (2021). Response of Global Tropical Cyclone Activity to Increasing CO2: Results from Downscaling CMIP6 Models. *Journal of Climate*, 34(1):57–70.

Entwisle, D. (2019). 3D geological modelling at the British Geological Survey (BGS). Publication Title: NTU-BCA Workshop on 3D Geological Modelling.

Etheridge, D. M., Steele, L. P., Langenfelds, R. L., Francey, R. J., Barnola, J.-M., and Morgan, V. I. (1996). Natural and anthropogenic changes in atmospheric CO_2 over the last 1000 years from air in Antarctic ice and firn. *Journal of Geophysical Research: Atmospheres*, 101(D2):4115–4128.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

Faulkner, D., Francis, O., and Lamb, R. (2012). Greenfield run off and flood estimation on small catchments: Small catchments' greenfield run off and flood estimation. *Journal of Flood Risk Management*, 5(1):81–90.

Few, R. (2003). Flooding, vulnerability and coping strategies: local responses to a global threat. *Progress in Development Studies*, 3(1):43–58.

Field, C.B.,, V. Barros,, T.F. Stocker,, D. Qin,, D.J. Dokken,, K.L. Ebi,, M.D. Mastrandrea,, K.J. Mach,, G.-K. Plattner,, S.K. Allen,, M. Tignor,, and P.M. Midgley (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Technical report, IPCC, Cambridge.

Fill, H. D. and Steiner, A. A. (2003). Estimating Instantaneous Peak Flow from Mean Daily Flow Data. *Journal of Hydrologic Engineering*, 8(6):365–369.

Foody, G., McCulloch, M. B., and Yates, W. B. (1995). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, 16(9):1707–1723.

Foody, G. M. and Arora, M. K. (1997). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, 18(4):799–810.

Forsee, W. J. and Ahmad, S. (2011). Evaluating Urban Storm-Water Infrastructure Design in Response to Projected Climate Change. *Journal of Hydrologic Engineering*, 16(11):865–873.

Fuller, R. M., Smith, G. M., Sanderson, J. M., Hill, R. A., and Thomson, A. G. (2002). The UK Land Cover Map 2000: Construction of a Parcel-Based Vector Map from Satellite Images. *The Cartographic Journal*, 39(1):15–25.

Fuller, W. E. (1914). Flood Flows. *Transactions of the American Society of Civil Engineers*, 77(1):564–617.

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., and Lin, Q. (2020). Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *Journal of Hydrology*, 589:125188.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. (2018). GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. M. A. (2018). Conditional Neural Processes. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

Gentry, A. H. and Lopez-Parodi, J. (1980). Deforestation and Increased Flooding of the Upper Amazon. *Science*, 210(4476):1354–1356.

Golmohammadi, G., Prasher, S., Madani, A., and Rudra, R. (2014). Evaluating Three Hydrological Distributed Watershed Models: MIKE-SHE, APEX, SWAT. *Hydrology*, 1(1):20–39.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts.

Goodfellow, I., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. (2014). Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. In *ICLR2014*.

Gordon, J., Bruinsma, W. P., Foong, A. Y. K., Requeima, J., Dubois, Y., and Turner, R. E. (2019). Convolutional Conditional Neural Processes. *International Conference on Learning Representations 2020*.

Govindaraju, R. S., Rao, A. R., and Singh, V. P., editors (2000). *Artificial Neural Networks in Hydrology*, volume 36 of *Water Science and Technology Library*. Springer Netherlands, Dordrecht.

Gray, D. M., editor (1973). *Handbook on the principles of hydrology: with special emphasis directed to Canadian conditions in the discussions, applications, and presentation of data*. Water Information Center, inc, Port Washington, N.Y.

Gray, W. M. (1979). Hurricanes: Their formation, structure and likely role in the tropical circulation. In *Meteorology Over the Tropical Oceans*, pages 155–218. Royal Meteorological Society, Bracknell.

Gray, W. M. (1998). The formation of tropical cyclones. *Meteorology and Atmospheric Physics*, 67(1-4):37–69.

Greenwood, G. (2020). We don't have the money to beat polluters, Environment Agency admits. *The Times*. Monday 28 December 2020.

Groenemeijer, P., Púčik, T., Holzer, A. M., Antonescu, B., Riemann-Campe, K., Schultz, D. M., Kühne, T., Feuerstein, B., Brooks, H. E., Doswell, C. A., Koppert, H.-J., and Sausen, R. (2017). Severe Convective Storms in Europe: Ten Years of Research and Education at the European Severe Storms Laboratory. *Bulletin of the American Meteorological Society*, 98(12):2641–2651.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91.

Gutjahr, O., Putrasahan, D., Lohmann, K., Jungclaus, J. H., von Storch, J.-S., Brüggemann, N., Haak, H., and Stössel, A. (2019). Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model Intercomparison Project (HighResMIP). *Geoscientific Model Development*, 12(7):3241–3281.

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX On the Matérn correlation family. *Biometrika*, 93(4):989–995.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J.-J., Mao, J., Mizielinski, M. S., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J., and von Storch, J.-S. (2016). High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geoscientific Model Development*, 9(11):4185–4208.

Haerter, J. O., Hagemann, S., Moseley, C., and Piani, C. (2011). Climate model bias correction and the role of timescales. *Hydrology and Earth System Sciences*, 15(3):1065–1079.

Haigh, I. D., Marcos, M., Dangendorf, S., and Mir Calafat, F. (2017). Editorial: Sea Level Variability and Change. *Frontiers in Marine Science*, 4.

Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., Kjeldsen, T. R., Kriaučiūnienė, J., Kundzewicz, Z. W., Lang, M., Llasat, M. C., Macdonald, N., McIntyre, N., Mediero, L., Merz, B., Merz, R., Molnar, P., Montanari, A., Neuhold, C., Parajka, J., Perdigão, R. A. P., Plavcová, L., Rogger, M., Salinas, J. L., Sauquet, E., Schär, C., Szolgay, J., Viglione, A., and Blöschl, G. (2014). Understanding flood regime changes in Europe: a state-of-the-art assessment. *Hydrology and Earth System Sciences*, 18(7):2735–2772.

He, K. and Sun, J. (2015). Convolutional neural networks at constrained time cost. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5353–5360, Boston, MA, USA. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

He, Q., Shen, Z., Wan, M., and Li, L. (2020). Precipitable Water Vapor Converted from GNSS-ZTD and ERA5 Datasets for the Monitoring of Tropical Cyclones. *IEEE Access*, 8:87275–87290.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian Processes for Big Data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 282–290, Arlington, Virginia, USA. AUAI Press.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable Variational Gaussian Process Classification. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA. PMLR.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J. (2018). ERA5 hourly data on single levels from 1979 to present.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Horn, M., Walsh, K., Zhao, M., Camargo, S. J., Scoccimarro, E., Murakami, H., Wang, H., Ballinger, A., Kumar, A., Shaevitz, D. A., Jonas, J. A., and Oouchi, K. (2014). Tracking Scheme Dependence of Simulated Tropical Cyclone Response to Idealized Climate Simulations. *Journal of Climate*, 27(24):9197–9213.

Hsu, T.-W., Shih, D.-S., and Chen, W.-J. (2015). Destructive Flooding Induced by Broken Embankments along Linbian Creek, Taiwan, during Typhoon Morakot. *Journal of Hydrologic Engineering*, 20(7):05014025.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

Hunt, K. M. R. and Menon, A. (2020). The 2018 Kerala floods: a climate change perspective. *Climate Dynamics*, 54(3-4):2433–2446.

Hunter, D., Yu, H., Pukish, III, M. S., Kolbusz, J., and Wilamowski, B. M. (2012). Selection of Proper Neural Network Sizes and Architectures—A Comparative Study. *IEEE Transactions on Industrial Informatics*, 8(2):228–240.

Institute of Hydrology (2008). *Flood Estimation Handbook*. Number 2 in Flood estimation handbook / Institute of Hydrology. Centre for Ecology and Hydrology, Wallingford.

International Monetary Fund (2022). World Economic Outlook Database. Database - https://www.imf.org/en/Publications/WEO/weo-database/2022/April.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456, Lille, France. JMLR.org.

Jaiswal, R. K., Ali, S., and Bharti, B. (2020). Comparative evaluation of conceptual and physical rainfall–runoff models. *Applied Water Science*, 10(1):48.

Jiao, D., Xu, N., Yang, F., and Xu, K. (2021). Evaluation of spatial-temporal variation performance of ERA5 precipitation data in China. *Scientific Reports*, 11(1):17956.

Johnson, J. (2021). Environment Agency -v- Southern Water Services Limited.

Jones, M. A., Hughes, A. G., Jackson, C. R., and Van Wonderen, J. J. (2012). Groundwater resource modelling for public water supply management in London. *Geological Society, London, Special Publications*, 364(1):99–111.

Jonkman, S. N., Maaskant, B., Boyd, E., and Levitan, M. L. (2009). Loss of Life Caused by the Flooding of New Orleans After Hurricane Katrina: Analysis of the Relationship Between Flood Characteristics and Mortality. *Risk Analysis*, 29(5):676–698.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200093.

Kendon, M. and McCarthy, M. (2015). The UK 's wet and stormy winter of 2013/2014. *Weather*, 70(2):40–47.

Kessler, H., Mathers, S., and Sobisch, H.-G. (2009). The capture and dissemination of integrated 3D geospatial knowledge at the British Geological Survey using GSI3D software and methodology. *Computers & Geosciences*, 35(6):1311–1321.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., and Kirschbaum, D. B. (2017). So, How Much of the Earth's Surface Is Covered by Rain Gauges? *Bulletin of the American Meteorological Society*, 98(1):69–78.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kitamoto, A. (2022). Digital Typhoon. Database - http://agora.ex.nii.ac.jp/digital-typhoon/.

Knutson, T. R., Sirutis, J. J., Garner, S. T., Vecchi, G. A., and Held, I. M. (2008). Simulated reduction in Atlantic hurricane frequency under twenty-first-century warming conditions. *Nature Geoscience*, 1(6):359–364.

Kolen, J. F. and Kremer, S. C., editors (2001). *A field guide to dynamical recurrent networks*. IEEE Press, New York.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59.

Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., Burrows, R. M., DelVecchia, A. G., Fritz, K. M., Shanafield, M., Burgin, A. J., Zimmer, M. A., Datry, T., Dodds, W. K., Jones, C. N., Mims, M. C., Franklin, C., Hammond, J. C., Zipper, S., Ward, A. S., Costigan, K. H., Beck, H. E., and Olden, J. D. (2022). Assessing placement bias of the global river gauge network. *Nature Sustainability*.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12):5089–5110.

Kristensen, P., Whalley, C., and Klancnik, K. (2018). *European waters assessment of status and pressures 2018*. European Environment Agency, Copenhagen. OCLC: 1080393556.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lagmay, A. M. F., Agaton, R. P., Bahala, M. A. C., Briones, J. B. L. T., Cabacaba, K. M. C., Caro, C. V. C., Dasallas, L. L., Gonzalo, L. A. L., Ladiero, C. N., Lapidez, J. P., Mungcal, M. T. F., Puno, J. V. R., Ramos, M. M. A. C., Santiago, J., Suarez, J. K., and Tablazon, J. P. (2015). Devastating storm surges of Typhoon Haiyan. *International Journal of Disaster Risk Reduction*, 11:1–12.

Landsea, C. W. and Cangialosi, J. P. (2018). Have We Reached the Limits of Predictability for Tropical Cyclone Track Forecasting? *Bulletin of the American Meteorological Society*, 99(11):2237–2243.

Laukkonen, J., Blanco, P. K., Lenhart, J., Keiner, M., Cavric, B., and Kinuthia-Njenga, C. (2009). Combining climate change adaptation and mitigation measures at the local level. *Habitat International*, 33(3):287–292.

Laville, S. and McIntyre, N. (2020). Water firms discharged raw sewage into England's rivers 200,000 times in 2019. *The Guardian*. Wednesday 1 Jul 2020.

Lawrence, N. D. (2019). Data Science and Digital Systems: The 3Ds of Machine Learning Systems Design. arXiv:1903.11241.

Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (2016). Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In Hua, G. and Jégou, H., editors, *Computer Vision – ECCV 2016 Workshops*, volume 9915, pages 47–54. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.

Lee, C.-Y., Camargo, S. J., Sobel, A. H., and Tippett, M. K. (2020). Statistical–Dynamical Downscaling Projections of Tropical Cyclone Activity in a Warming Climate: Two Diverging Genesis Scenarios. *Journal of Climate*, 33(11):4815–4834.

Lim, B. and Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200209.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2019a). When Gaussian Process Meets Big Data: A Review of Scalable GPs. arXiv.

Liu, M., Vecchi, G. A., Smith, J. A., and Knutson, T. R. (2019b). Causes of large projected increases in hurricane precipitation rates with global warming. *npj Climate and Atmospheric Science*, 2(1):38.

Loodin, N. (2020). Aral Sea: an environmental disaster in twentieth century in Central Asia. *Modeling Earth Systems and Environment*, 6(4):2495–2503.

Lorenz, E. N. (1969). Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of Atmospheric Sciences*, 26(4):636 – 646. Place: Boston MA, USA Publisher: American Meteorological Society.

Luu, L. N., Scussolini, P., Kew, S., Philip, S., Hariadi, M. H., Vautard, R., Van Mai, K., Van Vu, T., Truong, K. B., Otto, F., van der Schrier, G., van Aalst, M. K., and van Oldenborgh, G. J. (2021). Attribution of typhoon-induced torrential precipitation in Central Vietnam, October 2020. *Climatic Change*, 169(3-4):24.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK ; New York.

Maraun, D. (2016). Bias Correcting Climate Change Simulations - a Critical Review. *Current Climate Change Reports*, 2(4):211–220.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A., and Mearns, L. O. (2017). Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, 7(11):764–773.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3):RG3003.

Maraun, D. and Widmann, M. (2018). *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, 1 edition.

Marsh, T. (2008). A hydrological overview of the summer 2007 floods in England and Wales. *Weather*, 63(9):274–279.

Marsh, T. and Hannaford, J. (2007). *The summer 2007 floods in England & Wales: a hydrological appraisal*. Centre for Ecology & Hydrology, Wallingford. OCLC: 751663940.

Mattos, C. L. C., Dai, Z., Damianou, A. C., Forth, J., Barreto, G. A., and Lawrence, N. D. (2016). Recurrent Gaussian Processes. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Matérn, B. (1986). *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

McGregor, J. L. (1997). Regional climate modelling. *Meteorology and Atmospheric Physics*, 63(1-2):105–117.

McKay, D. I. A., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., and Lenton, T. M. (2022). Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science*, 377(6611).

Met Office Hadley Centre (2019). UKCP18 Global Projections on a 60km grid over the UK for 1900-2100. Database - http://catalogue.ceda.ac.uk/uuid/6a44fecc0f3842faaea53ab617dd2047.

Mhaskar, H., Liao, Q., and Poggio, T. (2017). When and Why Are Deep Networks Better Than Shallow Ones? *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Micklin, P. (2007). The Aral Sea Disaster. *Annual Review of Earth and Planetary Sciences*, 35(1):47–72.

Micklin, P. (2010). The past, present, and future Aral Sea. *Lakes & Reservoirs: Science, Policy and Management for Sustainable Use*, 15(3):193–213.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J. (2008). Stationarity Is Dead: Whither Water Management? *Science*, 319(5863):573–574.

Mitchell, T. M. (2013). *Machine learning*. McGraw-Hill series in Computer Science. McGraw-Hill, New York, 1st edition.

Muchan, K., Lewis, M., Hannaford, J., and Parry, S. (2015). The winter storms of 2013/2014 in the UK : hydrological responses and impacts. *Weather*, 70(2):55–61.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts London, England. OCLC: 812073227.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J. (2018). A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *Journal of Advances in Modeling Earth Systems*, 10(7):1383–1413.

Nagesh Kumar, D., Srinivasa Raju, K., and Sathish, T. (2004). River Flow Forecasting using Recurrent Neural Networks. *Water Resources Management*, 18(2):143–161.

Nash, J. and Sutcliffe, J. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290.

Nogueira, M. (2020). Inter-comparison of ERA-5, ERA-interim and GPCP rainfall over the last 40 years: Process-based analysis of systematic and random differences. *Journal of Hydrology*, 583:124632.

Ordnance Survey (2017). 1:50 000 Scale Colour Raster.

Ostby, F. P. (1999). Improved Accuracy in Severe Storm Forecasting by the Severe Local Storms Unit during the Last 25 Years: Then versus Now. *Weather and Forecasting*, 14(4):526 – 543. Place: Boston MA, USA Publisher: American Meteorological Society.

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., van Ruijven, B. J., van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M., and Solecki, W. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. *Global Environmental Change*, 42:169–180.

Pall, P., Allen, M. R., and Stone, D. A. (2007). Testing the Clausius–Clapeyron constraint on changes in extreme precipitation under CO2 warming. *Climate Dynamics*, 28(4):351–363.

Park, N. (2021). Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2020. Dataset, Office for National Statistics, London.

Parry, S., Marsh, T., and Kendon, M. (2013). 2012: from drought to floods in England and Wales. *Weather*, 68(10):268–274.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR. Issue: 3.

Peduzzi, P., Chatenoux, B., Dao, H., De Bono, A., Herold, C., Kossin, J., Mouton, F., and Nordbeck, O. (2012). Global trends in tropical cyclone risk. *Nature Climate Change*, 2(4):289–294.

Peel, M. C. and Blöschl, G. (2011). Hydrological modelling in a changing world. *Progress in Physical Geography: Earth and Environment*, 35(2):249–261.

Piani, C., Weedon, G., Best, M., Gomes, S., Viterbo, P., Hagemann, S., and Haerter, J. (2010). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *Journal of Hydrology*, 395(3-4):199–215.

Poff, N. L. (2002). Ecological response to and management of increased flooding caused by climate change. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 360(1796):1497–1510.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519.

Prein, A. F., Rasmussen, R. M., Ikeda, K., Liu, C., Clark, M. P., and Holland, G. J. (2017). The future intensification of hourly precipitation extremes. *Nature Climate Change*, 7(1):48–52.

Priestley, M. D. K., Ackerley, D., Catto, J. L., Hodges, K. I., McDonald, R. E., and Lee, R. W. (2020). An Overview of the Extratropical Storm Tracks in CMIP6 Historical Simulations. *Journal of Climate*, 33(15):6315–6343.

Qian, Q., Jia, X., and Lin, Y. (2022). Reduced Tropical Cyclone Genesis in the Future as Predicted by a Machine Learning Model. *Earth's Future*, 10(2).

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for Activation Functions. arXiv: 1710.05941.

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*, volume 3176, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg.

Raudkivi, A. J. (1979). *Hydrology: An Advanced Introduction to Hydrological Processes and Modelling*. Pergamon Press Ltd., Oxford, England, first edition. edition. Publication Title: Hydrology : an advanced introduction to hydrological processes and modelling.

Reed, K. A., Bacmeister, J. T., Rosenbloom, N. A., Wehner, M. F., Bates, S. C., Lauritzen, P. H., Truesdale, J. E., and Hannay, C. (2015). Impact of the dynamical core on the direct simulation of tropical cyclones in a high-resolution global model: DYNAMICAL CORE IMPACT ON TC ACTIVITY. *Geophysical Research Letters*, 42(9):3603–3608.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., and DMIP Participants, a. (2004). Overall distributed model intercomparison project results. *Journal of Hydrology*, 298(1-4):27–60.

Refsgaard, J. C., Storm, B., and Clausen, T. (2010). Système Hydrologique Europeén (SHE): review and perspectives after 30 years development in distributed physically-based hydrological modelling. *Hydrology Research*, 41(5):355–377.

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J. C., Kc, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Da Silva, L. A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J. C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., and Tavoni, M. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, 42:153–168.

Ribeiro, R. P. and Moniz, N. (2020). Imbalanced regression and extreme value prediction. *Machine Learning*, 109(9-10):1803–1835.

Ricke, K. L. and Caldeira, K. (2014). Maximum warming occurs about one decade after a carbon dioxide emission. *Environmental Research Letters*, 9(12):124002.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C., Senan, R., Zarzycki, C., and Ullrich, P. (2020a). Impact of Model Resolution on Tropical Cyclone Simulation Using the HighResMIP–PRIMAVERA Multimodel Ensemble. *Journal of Climate*, 33(7):2557–2583.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vannière, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L., Chauvin, F., Terray, L., Valcke, S., Moine, M., Putrasahan, D., Roberts, C. D., Senan, R., Zarzycki, C., Ullrich, P., Yamada, Y., Mizuta, R., Kodama, C., Fu, D., Zhang, Q., Danabasoglu, G., Rosenbloom, N., Wang, H., and Wu, L. (2020b). Projected Future Changes in Tropical Cyclones Using the CMIP6 HighResMIP Multimodel Ensemble. *Geophysical Research Letters*, 47(14).

Robson, A. and Reed, D. (2008). *Statistical procedures for flood frequency estimation*. Number 3 in Flood estimation handbook / Institute of Hydrology. Centre for Ecology and Hydrology, Wallingford.

Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. A. (2009). A safe operating space for humanity. *Nature*, 461(7263):472–475.

Rogers, R., Aberson, S., Aksoy, A., Annane, B., Black, M., Cione, J., Dorst, N., Dunion, J., Gamache, J., Goldenberg, S., Gopalakrishnan, S., Kaplan, J., Klotz, B., Lorsolo, S., Marks, F., Murillo, S., Powell, M., Reasor, P., Sellwood, K., Uhlhorn, E., Vukicevic, T., Zhang, J., and Zhang, X. (2013). NOAA'S Hurricane Intensity Forecasting Experiment: A Progress Report. *Bulletin of the American Meteorological Society*, 94(6):859–882.

Ruder, S. (2017). An overview of gradient descent optimization algorithms. arXiv:1609.04747.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Runyan, C. and D'Odorico, P. (2016). *Global Deforestation*. Cambridge University Press.

Ruppert, J. H., Wing, A. A., Tang, X., and Duran, E. L. (2020). The critical role of cloud–infrared radiation feedback in tropical cyclone development. *Proceedings of the National Academy of Sciences*, 117(45):27884–27892.

Rädler, A. T., Groenemeijer, P., Faust, E., and Sausen, R. (2018). Detecting Severe Weather Trends Using an Additive Regressive Convective Hazard Model (AR-CHaMo). *Journal of Applied Meteorology and Climatology*, 57(3):569–587.

Samuels, P., Huntington, S., Allsop, W., and Harrop, J. (2008). *Flood Risk Management: Research and Practice: Extended Abstracts Volume (332 pages) + full paper CD-ROM (1772 pages)*. CRC Press.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Scardi M (1996). Artificial neural networks as empirical models for estimating phytoplankton production. *Marine Ecology Progress Series*, 139:289–299.

Schiermeier, Q. (2013). Did climate change cause Typhoon Haiyan? *Nature*.

Schindler, A., Lidy, T., and Rauber, A. (2016). Comparing Shallow versus Deep Neural Network Architectures for Automatic Music Genre Classification. In *FMT*.

Servat, E. and Dezetter, A. (1991). Selection of calibration objective fonctions in the context of rainfall-ronoff modelling in a Sudanese savannah area. *Hydrological Sciences Journal*, 36(4):307–330.

Sildir, H., Aydin, E., and Kavzoglu, T. (2020). Design of Feedforward Neural Networks in the Classification of Hyperspectral Imagery Using Superstructural Optimization. *Remote Sensing*, 12(6):956.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, pages 1–14. Computational and Biological Learning Society.

Singh, V. P., editor (1995). *Computer models of watershed hydrology*. Water Resources Publications, Highlands Ranch, Colo, rev. ed edition.

Smith, M. B., Seo, D.-J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S. (2004). The distributed model intercomparison project (DMIP): motivation and experiment design. *Journal of Hydrology*, 298(1-4):4–26.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian Processes using Pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, pages 2951–2959, Red Hook, NY, USA. Curran Associates Inc. event-place: Lake Tahoe, Nevada.

Son, J.-H., Kim, H., Heo, K.-Y., Kwon, J.-I., Jeong, S.-H., Choi, J.-Y., Chun, J.-Y., Kwon, Y.-Y., and Choi, J.-W. (2022). Strategy for the Prediction of Typhoon Wind and Storm Surge Height Using the Parametric Typhoon Model: Case Study for Hinnamnor in 2022. *Atmosphere*, 14(1):82.

Spencer, P., Faulkner, D., Perkins, I., Lindsay, D., Dixon, G., Parkes, M., Lowe, A., Asadullah, A., Hearn, K., Gaffney, L., Parkes, A., and James, R. (2018). The floods of December 2015 in northern England: description of the events and possible implications for flood hydrology in the UK. *Hydrology Research*, 49(2):568–596.

Srinivasan, R., Ramanarayanan, T. S., Arnold, J. G., and Bednarz, S. T. (1998). LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART II: MODEL APPLICATION. *Journal of the American Water Resources Association*, 34(1):91–101.

Stedinger, J. R. and Griffis, V. W. (2011). Getting From Here to Where? Flood Frequency Analysis and Climate1: Getting From Here to Where? Flood Frequency Analysis and Climate. *JAWRA Journal of the American Water Resources Association*, 47(3):506–513.

Steinruecken, C., Smith, E., Janz, D., Lloyd, J., and Ghahramani, Z. (2019). The Automatic Statistician. In Hutter, F., Kotthoff, L., and Vanschoren, J., editors, *Automated Machine Learning*, pages 161–173. Springer International Publishing, Cham. Series Title: The Springer Series on Challenges in Machine Learning.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K. (2018). A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics*, 56(1):79–107.

Tabachnick, W. J. (2010). Challenges in predicting climate and environmental effects on vector-borne disease episystems in a changing world. *Journal of Experimental Biology*, 213(6):946–954.

Tabari, H., Paz, S. M., Buekenhout, D., and Willems, P. (2021). Comparison of statistical downscaling methods for climate change impact analysis on precipitation-driven drought. *Hydrology and Earth System Sciences*, 25(6):3493–3517.

Tabor, K. and Williams, J. W. (2010). Globally downscaled climate projections for assessing the conservation impacts of climate change. *Ecological Applications*, 20(2):554–565.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Tanoue, M., Hirabayashi, Y., and Ikeuchi, H. (2016). Global-scale river flood vulnerability in the last 50 years. *Scientific Reports*, 6(1):36021.

Tans, P. and Keeling, R. (2022). Trends in Atmospheric Carbon Dioxide - Mauna Loa CO2 Annual Mean Data.

Tarek, M., Brissette, F. P., and Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences*, 24(5):2527–2544.

Taszarek, M., Allen, J. T., Groenemeijer, P., Edwards, R., Brooks, H. E., Chmielewski, V., and Enno, S.-E. (2020). Severe Convective Storms across Europe and the United States. Part I: Climatology of Lightning, Large Hail, Severe Wind, and Tornadoes. *Journal of Climate*, 33(23):10239–10261.

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075.

Teutschbein, C. and Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, 456-457:12–29.

The Environment Agency (2007). Review of 2007 summer floods.

The Environment Agency (2022). Catchment Data Explorer. Database - https://environment.data.gov.uk/catchment-planning.

The Global Runoff Data Centre (2022). GRDC stations with monthly data, including data derived from daily data. Database - https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser#dataDownload/Home.

Titsias, M. (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

Transparency International (2022). Corruption Perceptions Index. Database - https://www.transparency.org/en/cpi/2021.

Trošelj, J. and Lee, H. S. (2021). Modelling typhoon-induced extreme river discharges: A case study of Typhoon Hagibis in Japan. *Journal of Hydrology: Regional Studies*, 34:100776.

Turner, R. E. (2017). Sparse Gaussian Process Approximations.

UK Centre for Ecology & Hydrology (2022). UK National River Flow Archive Data. Database - https://nrfa.ceh.ac.uk/data.

V. Masson-Delmotte, P. Zhai, A. Pirani, S.L., Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell,, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.) (2021). IPCC, 2021: Summary for Policymakers. Technical report, Cambridge University Press, Cambridge, United Kingdom.

van Vuuren, D. P., Kriegler, E., O'Neill, B. C., Ebi, K. L., Riahi, K., Carter, T. R., Edmonds, J., Hallegatte, S., Kram, T., Mathur, R., and Winkler, H. (2014). A new scenario framework for Climate Change Research: scenario matrix architecture. *Climatic Change*, 122(3):373–386.

Verbout, S. M., Brooks, H. E., Leslie, L. M., and Schultz, D. M. (2006). Evolution of the U.S. Tornado Database: 1954–2003. *Weather and Forecasting*, 21(1):86–93.

Vitart, F., Anderson, J. L., and Stern, W. F. (1997). Simulation of Interannual Variability of Tropical Storm Frequency in an Ensemble of GCM Integrations. *Journal of Climate*, 10(4):745 – 760. Place: Boston MA, USA Publisher: American Meteorological Society.

Vitart, F. and Stockdale, T. N. (2001). Seasonal Forecasting of Tropical Storms Using Coupled GCM Integrations. *Monthly Weather Review*, 129(10):2521 – 2537. Place: Boston MA, USA Publisher: American Meteorological Society.

Walsh, K. J., McBride, J. L., Klotzbach, P. J., Balachandran, S., Camargo, S. J., Holland, G., Knutson, T. R., Kossin, J. P., Lee, T., Sobel, A., and Sugi, M. (2016). Tropical cyclones and climate change. *WIREs Climate Change*, 7(1):65–89.

Wang, C., Zhang, L., Lee, S.-K., Wu, L., and Mechoso, C. R. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, 4(3):201–205.

Wang, S.-C. (2003). Artificial Neural Network. In *Interdisciplinary Computing in Java Programming*, pages 81–100. Springer US, Boston, MA.

Wash, C. H., Heikkinen, S. H., Liou, C.-S., and Nuss, W. A. (1990). A Rapid Cyclogenesis Event during GALE IOP 9. *Monthly Weather Review*, 118(2):375 – 391.

Waters, C. N., Zalasiewicz, J., Summerhayes, C., Barnosky, A. D., Poirier, C., Gałuszka, A., Cearreta, A., Edgeworth, M., Ellis, E. C., Ellis, M., Jeandel, C., Leinfelder, R., McNeill, J. R., Richter, D. d., Steffen, W., Syvitski, J., Vidas, D., Wagreich, M., Williams, M., Zhisheng, A., Grinevald, J., Odada, E., Oreskes, N., and Wolfe, A. P. (2016). The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science*, 351(6269).

Wehner, M. F., Bala, G., Duffy, P., Mirin, A. A., and Romano, R. (2010). Towards Direct Simulation of Future Tropical Cyclone Statistics in a High-Resolution Global Atmospheric Model. *Advances in Meteorology*, 2010:1–13.

Wehner, M. F., Reed, K. A., and Zarzycki, C. M. (2017). High-Resolution Multi-decadal Simulation of Tropical Cyclones. In Collins, J. M. and Walsh, K., editors, *Hurricanes and Climate Change*, pages 187–211. Springer International Publishing, Cham.

Westermayer, A. T., Groenemeijer, P., Pistotnik, G., Sausen, R., and Faust, E. (2017). Identification of favorable environments for thunderstorms in reanalysis data. *Meteorologische Zeitschrift*, 26(1):59–70.

Weyant, J. (2017). Some Contributions of Integrated Assessment Models of Global Climate Change. *Review of Environmental Economics and Policy*, 11(1):115–137.

Wilby, R. L. and Quinn, N. W. (2013). Reconstructing multi-decadal variations in fluvial flood risk using atmospheric circulation patterns. *Journal of Hydrology*, 487:109–121.

Willis, K. G. (1995). The Benefits of Alleviating Low Flows in Rivers. *International Journal of Water Resources Development*, 11(3):243–260.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The Marginal Value of Adaptive Gradient Methods in Machine Learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Woodruff, J. D., Irish, J. L., and Camargo, S. J. (2013). Coastal flooding by tropical cyclones and sea-level rise. *Nature*, 504(7478):44–52.

Working Party on the Hydraulic Design of Storm Sewers and Hydraulics Research Limited (1981). *Design and analysis of urban storm drainage: the Wallingford procedure. Volume 4. Volume 4.* National Water Council : Hydraulics Research Ltd., London; Wallingford. OCLC: 1221728231.

World Bank and United Nations (2010). *Natural Hazards, UnNatural Disasters: The Economics of Effective Prevention*. The World Bank.

Wright, D. B., Knutson, T. R., and Smith, J. A. (2015). Regional climate model projections of rainfall from U.S. landfalling tropical cyclones. *Climate Dynamics*, 45(11-12):3365–3379.

Wu, L., Wang, R., and Feng, X. (2018). Dominant Role of the Ocean Mixed Layer Depth in the Increased Proportion of Intense Typhoons During 1980–2015. *Earth's Future*, 6(11):1518–1527.

Yan, J., Mu, L., Wang, L., Ranjan, R., and Zomaya, A. Y. (2020). Temporal Convolutional Networks for the Advance Prediction of ENSO. *Scientific Reports*, 10(1):8055.

Yilmaz, K. K., Gupta, H. V., and Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model: Process-Based Diagnostic Evaluation of Hydrologic Model. *Water Resources Research*, 44(9).

Yoshida, K., Sugi, M., Mizuta, R., Murakami, H., and Ishii, M. (2017). Future Changes in Tropical Cyclone Activity in High-Resolution Large-Ensemble Simulations. *Geophysical Research Letters*, 44(19):9910–9917.

Zhao, M., Held, I. M., Lin, S.-J., and Vecchi, G. A. (2009). Simulations of Global Hurricane Climatology, Interannual Variability, and Response to Global Warming Using a 50-km Resolution GCM. *Journal of Climate*, 22(24):6653–6678.