Advancing crop disease early warning in South Asia by complementing expert surveys with internet media scraping

Jacob W. Smith[1], Asif Al Faisal[2], David Hodson[3], Suraj Baidya[4], Madan Bhatta[4], Dhruba Thapa[4], Roshan Basnet[5], William Thurston[6], Timothy J. Krupnik[2], and Christopher A. Gilligan[1]

[1] Department of Plant Sciences, University of Cambridge, Cambridge, UK.
[2] International Maize and Wheat Improvement Center (CIMMYT)-Bangladesh, Dhaka, Bangladesh.
[3] CIMMYT-Mexico, Texcoco, Mexico.
[4] Nepal Agricultural Research Council, Kathmandu, Nepal.
[5] National Wheat Research Program, Rupandehi, Nepal.
[6] Met Office, Fitzroy Road, Exeter, UK.

E-mail: jws52 at cam.ac.uk

Abstract

Wheat contributes one fifth of global food supply with an estimated 29% of global production in low and lower-middle income countries. As production expands across southern Asia, yields are often negatively impacted by outbreaks of fungal rust diseases. A wheat rust early warning and advisory system comprising surveillance, near real-time disease risk forecasts, and advisory dissemination has been established in two target countries in South Asia, including Nepal and Bangladesh. However, as wheat rust spores can be aerially transmitted over long distances, near real-time estimates of disease incidence are required from sources of infection in neighbouring regions. To address this challenge, we developed and tested a novel algorithm to generate proxy observations of infection sources using online media reports in two neighbouring countries, India and Pakistan. Media sampling could provide an effective alternative where data from ground surveys are not readily available in near real-time. Our results show that west Nepal was exposed to a substantial inoculum pressure from aerially-dispersed stripe rust spores originating from India and Pakistan. There were no outbreaks of stripe rust disease in Bangladesh with only very low levels of cross-border dispersion and generally unfavourable environmental conditions for infection. We further describe how proxy observations informed farmer decision-making in near real-time in Nepal and filled a knowledge gap in identifying early sources of infection for a major outbreak of stripe rust during 2020 in Nepal. Our results highlight the importance of international cooperation in mitigating transboundary plant pathogens.

Keywords: Early warning system, long-distance dispersal, disease surveillance, wheat stripe rust, Nepal, media scraping.

1. Introduction

The security of global food supply is threatened by conflict, economic shocks, rising inequality, climate change, and pandemics (Bentley et al., 2022). Almost half of the world's hungry are in South Asia (see Tables 1-4, FAO et al., 2022), which is also one of the largest wheat-producing regions, accounting for 50.6 M ha, 23 % of global wheat production in 2020 (FAOSTAT, 2023). Cultivation is concentrated along the Indo-Gangetic Plain, spanning northern regions of Pakistan and India to southern Nepal and Bangladesh. In this region wheat continues to be threatened by epidemics of airborne fungal pathogens, especially wheat rusts (Saari and Prescott, 1985; Bhavani et al., 2022), due to conducive climatic conditions (Kisana et al., 2003; Afzal et al., 2009; Ali et al., 2009, 2014). Urediniospores of rust pathogens (hereafter called spores) are capable of dispersing over long distances, increasing the likelihood of transmision across international territories and disrupting ongoing national control programmes in neigbouring countries (Meyer et al., 2017b; Radici et al., 2022).

Resistant varieties are routinely deployed to prevent infections from the dominant races of wheat rust species with many cases of long-term success. Epidemics still occur, however, when new races of the pathogen arise to which widely grown wheat varieties are not resistant (Singh et al., 2015; Bhavani et al., 2022). Effective control of the pathogen then depends upon predicting disease risk and early warning systems to enable farmers to apply fungicides in time to prevent yield loss.

An Early Warning and Advisory System (EWAS), originally developed and implemented for wheat rusts in Ethiopia (Allen-Sader et al., 2019), has been adapted and deployed in Nepal and Bangladesh (Bhavani et al., 2022), within South Asia. The EWAS involves field surveillance by trained experts (to identify new sources of infection), weather-driven dynamic models (to predicit sites at risk in the target countries from spore dispersal and environmental suitability for infection) and advisory reports (to communicate risk to growers) in near real-time. Early warnings are provided to farmers for the three most damaging wheat rusts: stripe, stem and leaf rust, caused, respectively, by *Puccinia striiformis* f. sp. *tritici* (*Pst*), *P. graminis* f. sp. *tritici*, and *P. triticina.* The advisory reports on infection levels, risk, and management options, are disseminated via extension agents and phone alerts, giving farmers up to three weeks in which to apply fungicides to mitigate the risk of infection and subsequent crop loss.

A major challenge in forcasting airborne plant pathogens is the limitation in obtaining timely surveillence data across the full area of influence (Carvajal-Yepes et al., 2019; Morris et al., 2022). The ability of wheat rust spores to disperse beyond international borders results in the potential for emerging epidemics to originate from neighbouring countries (Brown and Hovmøller, 2002), where near real-time surveillance may be unavailable. Focusing on Nepal and Bangladesh as target countries, a significant risk may arise from any occurrences of wheat rust to the west and south, in India and Pakistan. To the north, a barrier to transmission of viable spores is formed by the Himalayas. To the east, wheat production is relatively low.

Local and regional news outlets routinely publicise reports of wheat rust online. Web scraping is an established approach to extract online data into a structured format for analysis (Mitchell, 2015; Diouf et al., 2019) that has been applied in many studies of human infectious disease surveillance (Jahanbin et al., 2019; Pilipiec et al., 2023), as well as environmental research (Ghermandi and Sinclair, 2019) including pest monitoring (Daume, 2016). However, online news reports have not yet been tested as a complement to expert surveillance of crop diseases.

In the current study we undertake a media-scraping exercise using openly available software to generate proxy surveys of wheat rust infection in near real-time by web-scraping online news outlets for relevant reports beyond the target regions. Specifically, we compare two approaches (manual and automated scraping) to extract information from online news outlets covering India and Pakistan. Using historic weather data as driving variables and the Lagrangian spore dispersal models used in the EWAS, we calculate time series, source strengths and mapped densities of spore deposition over Nepal and Bangladesh from the scraped media sources. By comparing these with the corresponding outputs from sources identified by field surveys in Nepal, we show that there is strong evidence for the involvement of external sources in accounting for a *Pst* epidemic in Nepal in 2020 connected with the presence of multiple virulent pathotypes, including 238S119 previously seen in India but not in Nepal (Baidya et al., 2022). Finally, we present the results of operational forecasts of the EWAS that were conducted in near real-time from 2020 onwards to assess the benefit of accounting for external sources of infection to provide an effective early warning system for wheat rusts in Nepal and Bangladesh.

2. Data and Methods

2.1 Surveillance for wheat rust in Nepal and Bangladesh

Routine surveillance was conducted by trained pathologists throughout wheat-growing areas in Nepal and Bangladesh during 2020-2023. Surveyors monitored for all three species of wheat rust. The majority of surveys were conducted between January and March when the main season crop was maturing and was most vulnerable to infection. Taking a transect approach in pre-determined districts being visited by car, sites were generally randomly selected along the roadside approximately every 10 km. At each site, surveyors recorded the date, location coordinates, field area, wheat growth stage, and wheat cultivar (if known) as well as disease type, severity (measure of disease level on wheat plants), incidence (fraction of surveyed field that appears infected) and host's reaction type (Ali and Hodson, 2017). Expert surveys are recorded electronically on an OpenDataKit (ODK, 2023) survey form and gathered in near real-time on an ODK database for automated provision to Early Warning and Advisory System forecast models, and later uploaded to the global Rusttracker repository (RustTracker, 2023).

The levels of stripe rust disease varied amongst years with a potentially severe epidemic evident during 2020 in Nepal and comparatively low levels of disease in other years. There were no outbreaks of stripe rust disease in Bangladesh during the period of interest. Analysis of the effects of external inoculum sources estimated by media scraping were therefore carried out for the Nepal epidemic of stripe rust in 2020.

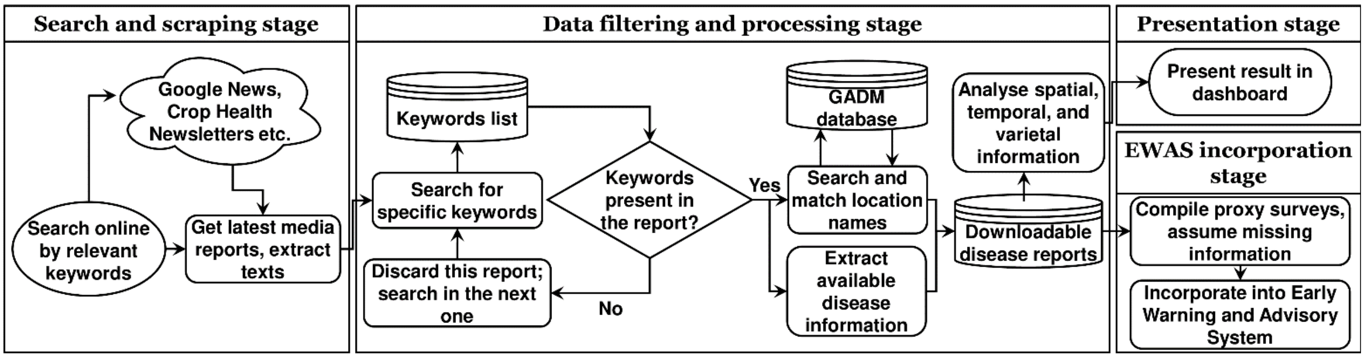

Figure 1: Workflow of the scraper tool for media reports of wheat rust infection (for more details, see text and Faisal, 2023b).

2.2 Scraped web media to estimate inoculum sources from non-target countries

To construct proxy surveys of wheat rust infection outside Nepal and Bangladesh from online media, we identified relevant news reports on rust disease and extracted information on observation date, location, area affected, crop growth stage, rust species, incidence and severity for each year from 2020 to 2022. Two methods of data collection were compared: one involving regular 'manual' searches, the other used automated computer searches of media reports. Data from each method were used to test the likely impact of wheat rust dispersed from these sources on the infection risk in Nepal and Bangladesh.

2.2.1 Manual search for news reports and compilation of proxy surveys

Weekly internet web searches were conducted during the susceptible period of the main wheat season (January to March) in 2020 for news reports in India or Pakistan that included the terms 'wheat rust', 'wheat yellow rust', 'yellow rust' and 'leaf rust'. Early investigation indicated that a large number of the major newspaper outlets in Pakistan and India provided English translations that were accessible by the news search API, and therefore the coverage was considered sufficient. Surveillance data from the December 2019 newsletter of the Indian Institute of Wheat and Barley Research (ICAR - IIWBR, 2020b) were included in the manual search because these records preceded the period of susceptibility to wheat rust diseases in Nepal during 2020.

Search results were filtered for relevance and key information extracted manually (an example is provided in supplementary Figure S1). The location coordinates for reports for all three (stripe, stem or leaf) rusts were identified as precisely as possible, sometimes reaching village level, and collated to provide spatially and temporally resolved proxy rust surveys. Detailed information on the level of infection and area diseased was missing in almost all cases. Therefore, the following default values were assumed: 1 hectare for affected crop area with medium disease incidence (20-40%) and medium severity from which to calculate source strengths, in line with characterisations of field surveys used by Allen-Sader et al. (2019). The wheat growth stage was estimated from the report date (Table 1). We assumed medium levels of incidence and severity since we anticipated low levels are more likely to remain unobserved, and high levels of disease may be observed earlier as medium levels. If a different disease level and observed area were assumed, because it would be applied to all proxy observations in the region, the impact on results would be spatially uniform (i.e., timing of influence is unaffected). Subsequent analyses focused on stripe rust during 2020 to assess the impact of external sources of inoculum on the epidemic in Nepal.

Table 1: Wheat growth stage assumptions for proxy surveys where information was unavailable following extraction from scraped media reports.

Report date within range:	Assumed wheat growth stage stage
24th Dec –	Tillering
18th Feb –	Booting
28th Feb –	Heading
9th Mar –	Flowering
14th Mar –	Milking
26th Mar –	Dough
10th Apr –	Maturity
20th Apr – 24th Dec	Outside of main season for winter wheat

2.2.2 Automated search for news reports and compilation of proxy surveys

Seeking a more efficient method for media scraping, we set up an automated identification system using Python (Figure 1; see Faisal, (2023b), for more details). The process starts with scheduled searching of multiple web domains for local and regional news outlets using the Google Custom Search Application Programming Interface (API). We used a primary search using the following English key words ("wheat rust attack India", "yellow rust spotted India", "wheat rust attack Pakistan", and "yellow rust spotted Pakistan") followed by a secondary filter against a pre-populated list of keywords (Faisal, 2023b). Reports were then processed to extract available disease information including report date, location names, affected site area, and cultivar names. Location coordinates were obtained by cross-referencing location names with the GADM database of locations associated with each administrative district of Pakistan and India. The GADM database was used because of the low computational demand required.

The processed infection reports were summarised on an online dashboard (Faisal, 2023a). The dashboard provided users with maps, time series, and distributions of disease prevalence, affected varieties and affected administrative districts based upon media available in Pakistan and India. The infection reports were extracted from the dashboard via an application programming interface (API), and a manual quality check was conducted to discard any remaining irrelevant news reports; for example, reports that provided a general warning to farmers but did not describe a specific outbreak. Proxy surveys were compiled from the autmatically extracted news reports with identical assumptions about source strength and crop growth stage as for manually extracted media reports.

2.3 Wheat rust source calculation

Spore dispersal simulations require the identification of source terms. For retrospective analysis of the 2020 main wheat season, three spore source terms were calculated: one based on known sites of wheat stripe rust infection from expert surveys for Nepal, one based on proxy surveys from manual-scraped media for Pakistan and India, and another based on proxy surveys from automated scraped media for Pakistan and India. Source terms were estimated using the method of Allen-Sader et al. (2019). The disease prevalence (incidence and severity) for each reported survey was scaled to give a spore emission per unit area per day (in the range 10^{11}–10^{13} spores ha^{-1} day^{-1}). The duration each survey was assumed to remain active (i.e. informing calculations of spore availability for passive release and dispersal) was based on the reported growth stage and the estimated days until senescence.

The full set of surveys was clustered according to administrative districts. The source location was defined as the site with highest prevalence in the district. Daily spore production (also referred to as source strength) was calculated from the area-weighted average of all active surveyed areas and scaled-up by the wheat area for a given district. MapSPAM2005 was used to apportion wheat production areas because of its comprehensive geographical coverage and similar resolution to the meteorological model (SPAM2005; IFPRI and IIASA, 2015). The more recent MapSPAM2010 was not used because of expert knowledge identifying inaccuracies in the area being investigated.

2.4 Wheat rust spore dispersal model: retrospective analysis

We simulated *Pst* spore dispersal from each of the three spore source terms in order to assess the impact of external sources identified by proxy surveys, with a focus on the epidemic of stripe rust in Nepal during 2020. The passive release, transport, spread, in-air viability, and deposition of wheat rust spores were calculated with the NAME dispersion model (Jones et al., 2007) modified to simulate wheat rust spores (Meyer et al., 2017a; b). All three spore dispersal simulations used the analysis meteorology dataset (i.e., the best estimate of the historical state of the atmosphere by assimilating available observations in a numerical weather predicition model) of the Unified Model with a resolution of three hours and a spatial resolution of approximately 0.14° longitude × 0.09° latitude (roughly 14 km × 10 km over South Asia) (Met Office, 2013). The principal output variable of interest was the number of viable spores deposited per unit area per day.

2.5 Wheat rust spore dispersal model: analysis of operational forecasts

Near real-time forecasting of spore deposition based on expert and proxy surveys was performed daily from 6th February 2020 as part of the wheat rust Early Warning and Advisory System (EWAS) in Bangladesh and Nepal. Using the 7-day global forecast from the UK Met Office Numerical Weather Prediction model (Walters et al., 2019), spore deposition was forecast with a resolution of three hours and spatial resolution of approximately 0.14° longitude × 0.09° latitude. In 2020 manual-scraped media data were used to provide near real-time identification of out-of-country proxy surveys in the wheat rust forecast EWAS for Nepal and Bangladesh. From 2021 onwards, data from the automated approach were used. The operational EWAS did not use manual and automated approaches in parallel to avoid duplication of surveys. Automatically identified news reports were extracted from the media scraper dashboard via an API and compiled as proxy surveys automatically each week. Following a manual quality check, relevant proxy surveys were provided to the EWAS to advise on wheat rust disease risk in Bangladesh and Nepal. In this study we investigate the impact of proxy survey information on near real-time forecasts of wheat rust risk.

3. Results

3.1 Expert surveys and status of wheat stripe rust during 2020 epidemic in Nepal

A total of 412 field surveys were conducted in Nepal between 1st February 2020 and 31st March 2020 across the main and summer season wheat growing areas in 45 districts across the seven provinces of the country. In the central and eastern mid-hills (areas above 250 m altitude), stripe rust was observed in 66% of surveys during February (Figure 2a). Prevalence receded in March when only 34% of surveys in the central and eastern mid-hills recorded the presence of stripe rust (Figure 2b). Almost no stripe rust was observed in the lowlands (below 250 m) – *terai* – of central and eastern Nepal.

A different infection pattern was observed in the west of Nepal. In March 2020, the most substantial outbreak of wheat stripe rust since 2005-2007 was recorded across the lowlands and mid-hills of west Nepal (Borlaug Global Rust Initiative (BGRI), 2020), when 60% of surveys at lowland sites and 89% of surveys at mid-hill sites reported stripe rust (Figure 2b). No stripe rust was observed in west Nepal in February 2020 (Figure 2a), although its presence in the western mid-hills was likely and cannot be ruled out. Pathotype analysis indicated the first appearance of a virulent *Pst* pathotype 238S119 in western areas of Nepal during the 2020 season (Baidya et al., 2022), which was also the dominant strain of the *Pst* pathogen in India at the time (ICAR - IIWBR, 2020a, 2021).

In Bangladesh during the 2020 wheat growing season, more than 2800 surveys were conducted and there were no reports of stripe rust infection.

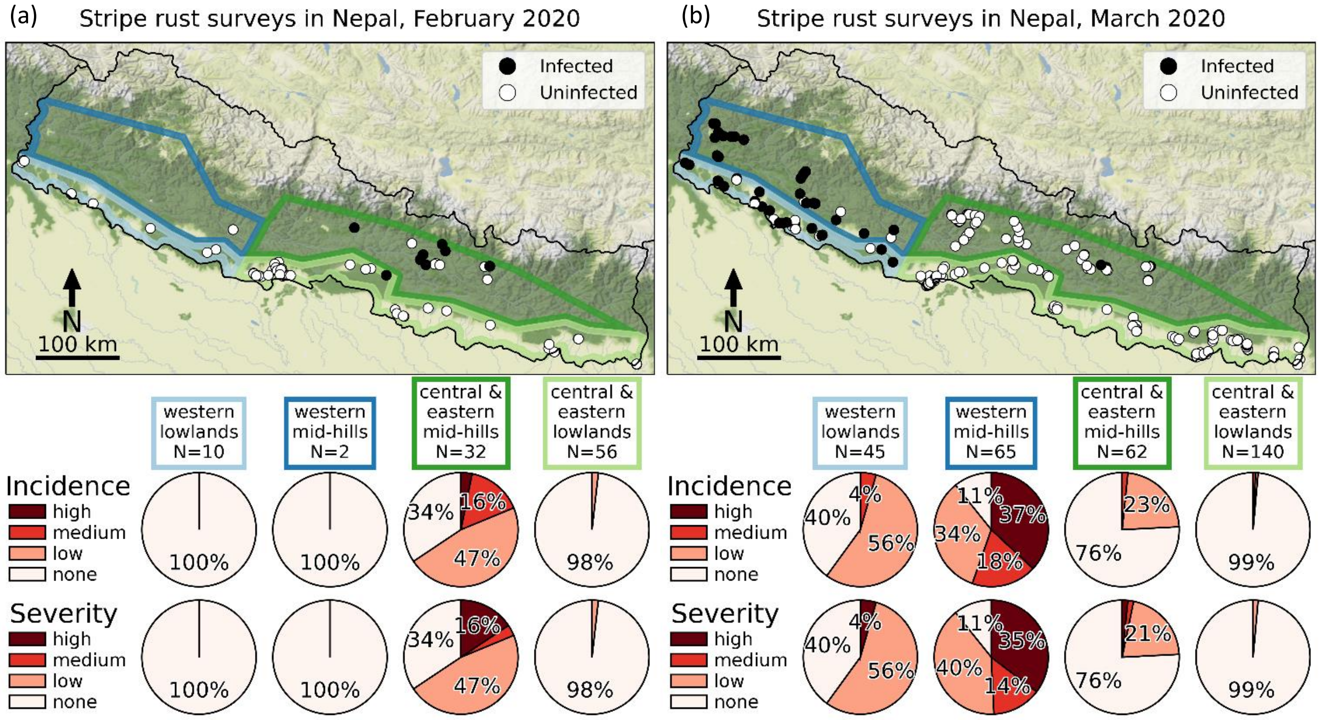

Figure 2: Locations and prevalence of stripe rust in expert surveys during (left) February and (right) March 2020. Bounded regions indicate western lowlands (light blue), western mid-hills (dark blue), central and eastern mid-hills (dark green) and central and eastern lowlands (light green). Surveys in Nepal lowlands occurred below an altitude of 250 m.

3.2 Comparison of manual and automated scraped media reports

3.2.1 Locations of proxy surveys from scraped media reports

Online news outlets reported the occurrence of stripe rust in wheat fields in Pakistan and northern India during the 2020 main season. The manual synthesis of media reports identified a total of 36 infection sites from 14 news reports spanning 14th January to 11th March 2020, whereas the automated search identified 43 infection sites from 15 news reports spanning 20th January to 31st March (and one additional site on 24th May).

Of these findings, four news reports were found by both methods and 26 of the manually identified infection sites corresponded to 28 of the automatically identified infection sites (for details, see Supplementary Information). The number of matching news reports was affected negatively by some cases of multiple news sites reporting the same occurrence of yellow rust, where the two media scraper methods have the potential to have made alternative decisions to retain and discard duplicates. From the December 2019 newsletter of the IIWBR, the manual compilation of proxy surveys identified three sites between 19th December 2019 and 29th January 2020.

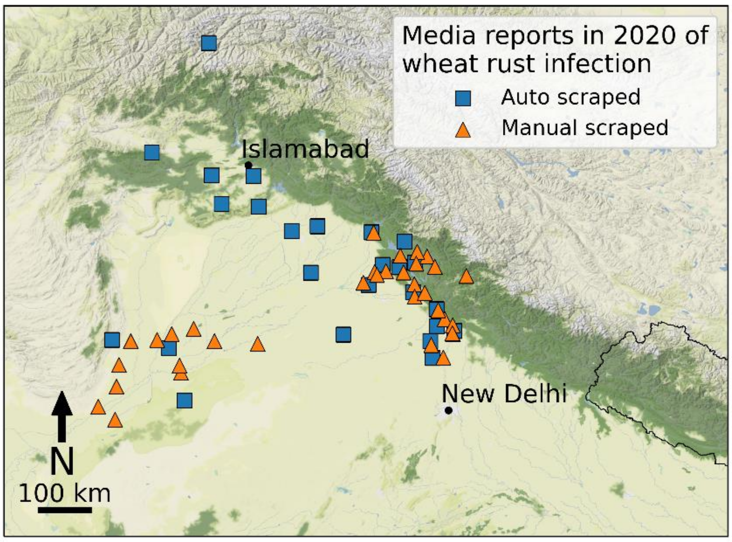

Figure 3: Sites of stripe rust infection in January-March 2020 identified by manual and automated media scraping.

Reported infection source locations were similar for both automated and manual methods at the national scale (within roughly 10 km of each other in most cases), albeit with the automated method retrieving a wider distribution of reports across the northern hilly areas of Pakistan (Figure 3) late in the season. By contrast, at a finer scale (<10 km), the reported site locations differed between the two methods, due to differences in identified news reports, as well as differences in the methods to extract location names and position them, as described in Section 2.2.

3.2.3 Calculation of source strength and spore deposition from proxy surveys

The calculations of source strengths from scraped web media show substantial quantities of *Pst* spores available for release (Figure 4a). Daily source strengths based on expert surveys in Nepal alone indicate a maximum of ~10^{17} spores day^{-1}. Sources in India and Pakistan were found to have at least ten times more spores available for release per day. Differences in source strength between the expert surveys and scraped media methods primarily relate to the regional estimates of wheat production. Sources also appeared at least 23 days earlier from proxy surveys in India and Pakistan during 2020 than from sites based on expert surveys in Nepal (Figure 4a).

Both the manual and automated methods result in a similar incremental increase in infection source strengths between January and February as new media reports of infection were published and incorporated (Figure 4a). However, the manual method identified larger source strengths, in part due to additional inclusion of proxy surveys scraped from the IIWBR newsletter for December 2019. While most spores are deposited locally, calculations with historical meteorology show the impacts of dispersal from scraped media sources extending many hundreds of kilometres, as far as Nepal and Bangladesh (Figure 4b-d). This pattern is consistent with single-source dispersal calculations across 2003-2014 by Meyer et al. (2017b). The spatial pattern of spore deposition is similar from both scraped media methods, with the highest levels of deposition around central Pakistan and across the Indo-Gangetic plain and foothills south of the Himalayan mountain range.

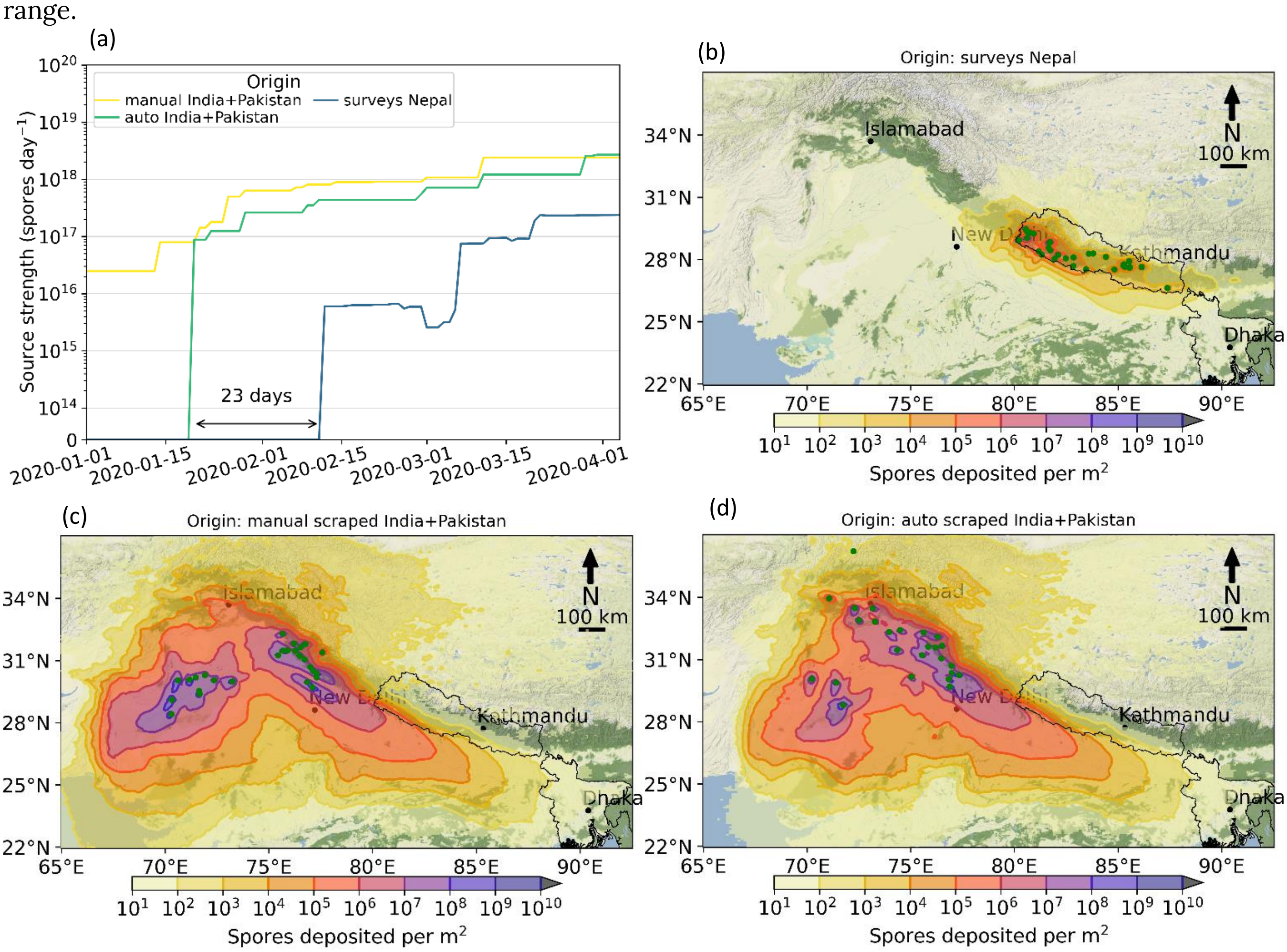

Figure 4: Time series and maps of simulated Pst spores during the 2020 growing season (1st Jan to 4th April) with different methods: (a) time series of regional spore availability from source terms, and (b,c,d) maps of source locations (green points) and number of spores deposited. Green points indicate source locations based on clustered surveys, see section 2.3 for details.

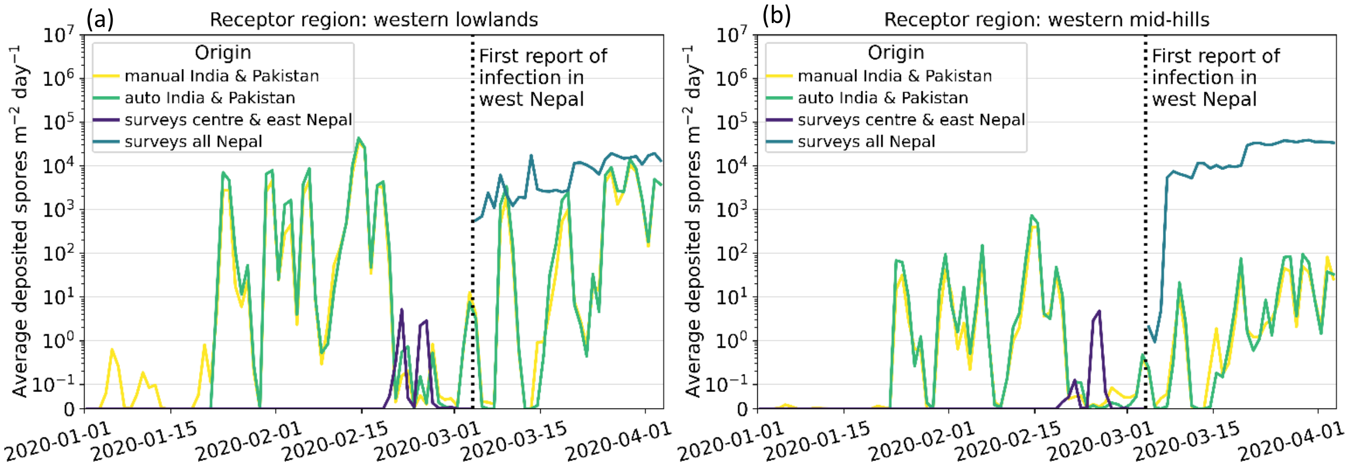

Figure 5: Simulated Pst spore deposition amounts per day in 2020 across (a) Nepal western lowlands and (b) Nepal western mid-hills from different source regions. The dotted line indicates the date of the first observation of stripe rust in west Nepal. The same time series analysis for receptor regions in central and eastern Nepal are shown in supplementary Figure S2a-b.

3.4 Impact of external inoculum pressure on Nepal and Bangladesh: retrospective analysis

The time series for *Pst* spore deposition in west Nepal simulated by the spore dispersal model from sources within and beyond Nepal are shown in Figure 5 (see also SI Fig S1). There were no recorded sources of *Pst* in Bangladesh. Results for both the manual and automatic media scraping methods were similar (the Spearman rank correlation for spore deposition over Nepal western lowlands is 0.94, and 0.96 for western mid-hills. For further analysis, see Supplementary Information), indicating that differences in the methods of media scraping are small relative to the impact of local meteorological conditions on long-distance dispersal into Nepal from different release sites.

The external sources from India and Pakistan contributed an additional 16% (manual) and 22% (automatic) load of *Pst* spores in Nepal compared with in-country sources, over the entire study period. The earlier occurrence of deposition in Nepal from out-of-country proxy surveys than from in-country expert surveys (Figure 5) reflects earlier infection of sites in India and Pakistan (Figure 4a).

We recall that there were no reported cases of stripe rust in west Nepal prior to March 2020 when an outbreak occurred that included a virulent pathotype 238S119 previously unseen in Nepal (Baidya et al., 2022) but known to be present in India (ICAR - IIWBR, 2020a, 2021). Stripe rust was reported at low levels in central and eastern Nepal in February and subsided for the rest of the season (Figure 2b). However, dispersal calculations do not support transmission of *Pst* from central and eastern Nepal to the west as the average spore deposition in western areas originating from central and eastern areas did not exceed 5 spores m^{-2} day^{-1} (Figure 5 purple lines). Stripe rust infections were reported in online news media in northern India and Pakistan, between January and March 2020 (Figure 3). Model simulations indicate *Pst* spores were present in northern India and Pakistan at least 23 days before stripe rust was detected in Nepal (Figure 4a), from which our dispersal simulations indicate suitable meteorological conditions for frequent deposition of *Pst* spores in western Nepal between 22nd January and 19th February, 2020 (Figure 5). That is three to seven weeks before the first infection reports in Nepal's western areas and long enough for infected fields to be detectable. Calculated deposition rates peaked at roughly 4×10^4 spores m^{-2} day^{-1} on the western lowlands and 7×10^2 spores m^{-2} day^{-1} on the western mid-hills (Figure 5 green lines) from outside Nepal with similar rates calculated locally in west Nepal after stripe rust became established (Figure 5 blue line). There was no evidence for involvement of Bangladesh as a source of *Pst* infection for Nepal during the 2020 main wheat-growing season.

Simulated *Pst* spore deposition over Bangladesh from reported infections in Nepal, India and Pakistan did not exceed 30 spores m^{-2} day^{-1} (see supplementary Figure S2c), indicating the *Pst* dispersal connection was weak. The low *Pst* inoculum pressure in simulations is consistent with the lack of stripe rust reports from surveys in Bangladesh, however environmental conditions are generally unsuitable for *Pst* in Bangladesh and therefore limit the chance of stripe rust infection.

3.5 Impact of external inoculum pressure on Nepal: near real-time forecast modelling

The above results used analysis (i.e. historic) weather data to enable a retrospective assessment of the role of alternative external sources of inoculum on the Nepal epidemic of *Pst* during 2020. We now assess the impact of external sources of inoculum on near real-time risk modelling using 7-day forecast weather data conducted for Nepal during the 2020 season when media reports were scraped manually.

A series of the seven-day forecasts for *Pst* spore deposition in Nepal is shown in Figure 6 for spore dispersal data from sources in Nepal (Figure 6b,d,e) and from proxy sources in Pakistan and India (Figure 6a,c,e. See also supplementary Figure S3 and supplementary Videos 1 and 2). Forecasts indicated an early and persistent influence of *Pst* spores from beyond the borders of west Nepal (Figure 6a) and that infections recorded in central and eastern Nepal did not provide substantial inoculum pressure over west Nepal (Figure 6d). Allowance for sources of external of inoculum derived from the scraped media analysis enabled early warnings and advice to be communicated to farmers through extension agencies in Nepal for farmers to apply fungicide to mitigate the risk of wheat stripe rust infection.

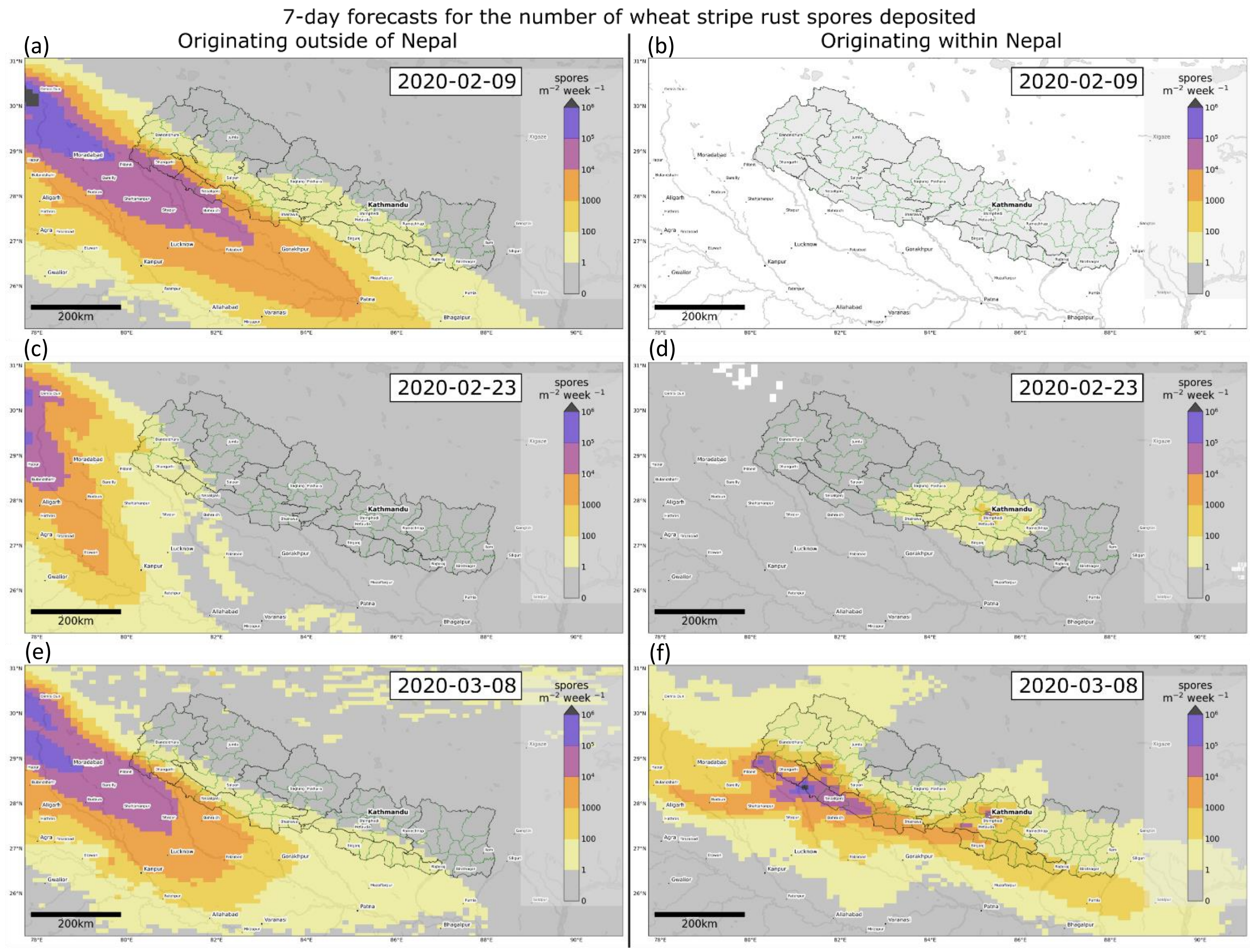

Figure 6: Examples of seven-day risk forecasts of Pst *spore deposition during the 2020 season based on near real-time information from (a,c,e) manually scraped online news media, and (b,d,e) expert pathologist surveys within Nepal on (a,b) 9th February, (c,d) 23rd February, and (e,f) 8th March. A video of weekly seven-day risk forecasts is shown in supplementary* Video 1. *The forecast of combined deposition from both scraped media and surveys is shown in supplementary Figure S3, and in supplementary* Video 2.

4. Discussion and conclusions

Our primary aim was to assess how scraped media reports of wheat rust infection could be used as a novel proxy for field surveys in non-target countries. While the manual and automated media scraping searches came up with different site locations within India and Pakistan, the exact location of the spore sources was less important for dispersal over long distances as their overall effects were similar on potential spore dispersal and risk of deposition in Nepal and Bangladesh. Spore dispersal calculations show connectivity of stripe rust occurrences in neighbouring countries with Nepal. Sources *outside* Nepal were calculated to account for an additional 16% to 22% inoculum pressure *within* Nepal (for manual and automated methods of media scraping, respectively). Our result indicates the importance of allowing for potential sources of long distance dispersal in wheat rust early warning systems, previously identified by Meyer et al. (2017b).

We investigated a possible precursor to the sudden outbreak of stripe rust in west Nepal during 2020 and found long-distance dispersal from stripe rust occurrences in India and Pakistan to be a possible contributor, in agreement with the first appearance of a virulent strain of *Pst* in west Nepal (Baidya et al., 2022). Dispersal calculations based on near real-time field surveillance by trained personnel in central and eastern Nepal suggested no causal connection with earlier infections of stripe rust in the rest of Nepal (prior to stripe rust arriving in west Nepal, simulations indicate the number of viable deposited spores originating from outside Nepal exceeded those from central and eastern Nepal by a factor of around 8400).

The outbreak in west Nepal developed suddenly, indicating the potential emergence or incursion of a new virulent race (see, for example, Chen, 2020) rather than carryover from earlier crops in the same area. Barberry is a documented functional alternate host for *Pst* (Jin et al., 2010), a potential source of early season infections, and a source of new pathogen diversity through sexual reproduction (e.g., Mehmood et al., 2020). Several studies indicate a potential role for barberry in Nepal (Khan et al., 2019; Hovmøller et al., 2023) but conclusive evidence is lacking and further research is needed. A role for barberry in the 2020 main season *Pst* development cannot be ruled out, however long-distance dispersal of spores from external sources, including a new virulent strain, appears to have also contributed to the outbreak.We note that disease control is not accounted for in the spore source calculations as reliable data about fungicide use are not available. As a result, simulations overestimate inoculum pressure and therefore represent a worst case scenario.

Complementing expert surveys with scraped web media has informed in-season advisories disseminated through extension services to farmers in Nepal and Bangladesh since February 2020 and, in particular, provided advance warning of the substantial stripe rust outbreak that occurred in Nepal during the 2020 main season (ARRCC, 2022). Cooperation of surveillance between neighbouring nations is key in managing transboundary plant pathogens (Thompson et al., 2016; Jansen and De La Cruz Bekema, 2023; Radici et al., 2023) and has been a noted success of many multi-national efforts (e.g., Bhavani et al., 2022; Global Rust Reference Centre, 2024). Near real-time field surveillance offers the most accurate view of disease status but is costly and depends on well-coordinated reporting systems that can ideally be integrated across national boundaries. Proxy surveys from scraped news media are a novel data source for plant disease monitoring. They have the potential for low-cost high-coverage rapid application in disease early warning systems.

The validity of online news reports in India and Pakistan as a proxy for expert field-based surveillance observations was inferred by their attribution to timing of observed disease in Nepal that could not have arisen from sources in Nepal because of prevailing wind conditions. A more rigorous test involving comparison of field surveillance with media-scraped data for the same region and season was not possible because of the unavailability of field-based surveillance data from the media-assessed countries in this study, but is indeed crucial for future assessment. Moreover, formal validation is further complicated in that media-sourced and field survey data are not necessarily independent. Detailed inspection of the media-scraped data confirms that media reports frequently cite field survey reports as supporting evidence (for example, see supplementary Figure S1; for a full listing see the data availability statement).

It remains important that early warning systems consider different sources of information separately. Web scraping poses many of the same challenges as data gathered from social media, namely noise, bias, and future availability (Ghermandi and Sinclair, 2019). In the case of this study, the representativeness of news reports is subject to the resources and interests of each media outlet. For instance, news media are unlikely to report on the absence of rusts. Indeed, in the 2021 to 2023 main seasons, stripe rust was relatively limited in India and Pakistan, and the automated media scraper identified relatively few reports relating to wheat rust (5, 13 and 0 reports, respectively: the stripe rust forecasts for 2021 and 2022 are shown in supplementary Videos 3 and 4, respectively.) News reports of wheat rust presence may be relatively more common in India and Pakistan than in many other countries since agriculture is a major part of the national economies (accounting for 16.8% and 22.7% of national GDP in 2021, respectively; World Bank, 2021) and national wheat institutes engage with news outlets (as demonstrated by the proportion of identified news reports of wheat rust occurrences that quote plant pathologists).Past studies provide a number of approaches to tackle the general challenges of scraped media, including noise, bias, and future availability (Alomar et al., 2016; Daume, 2016; Ghermandi and Sinclair, 2019). Approaches that may enhance the novel integration with crop disease models presented in this paper include a direct comparison of proxy and expert surveys in the same region and season, multilingual functionality, fuzzy logic to improve location name identification, and the use of a more open web search API.

Our study has demonstrated a viable means of monitoring for wheat rust occurrence where near real-time surveillance is unavailable but public news outlets are engaged, offering a novel advance in applied epidemiological modelling to support plant health initiatives. Digital agriculture tools may continue to provide opportunities to share knowledge and enhance crop disease early warning systems to promote international co-operation in managing transboundary pathogens.

Data availability statement

The code supporting the media scraper tool and dashboard are available on a public global repository from Faisal (2023b). The code and data supporting the analysis for this paper, namely wheat rust surveys in Nepal during 2020, proxy surveys from scraped media, spore source calculations, spore deposition simulation results, analysis, and figure generation, are available on a public global repository (Smith, 2023). Additional data and supporting information for wheat rust surveillance, including surveys covering other periods and countries, is available online and on request from RustTracker (2023).

Funding statement

This work was funded by a grant from the Bill & Melinda Gates Foundation to the Cambridge group, as well as the United Kingdom Foreign, Commonwealth and Development Office's programme Asia Regional Resilience in a Changing Climate (ARRCC). This work is also mapped to the One CGIAR Regional Integrated initiative Transforming Agrifood Systems in South Asia (TAFSSA, 2023). Accordingly, we would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund: https://www.cgiar.org/funders/.

Acknowledgements

We thank the collaborators of the ARRCC Early Warning and Advisory System project who have guided it to useful outcomes, including Moin Salam, Washiq Faisal, Tamás Mona, Sarah Millington, and George Gibson. We also thank the anonymous reviewers whose comments improved the paper, as well as Alison Scott-Brown for helpful discussions, and Lawrence Bower for technical support. Map tiles (C) Stadia Maps (C) Stamen Design (C) OpenMapTiles (C) OpenStreetMap contributors, licensed under CC BY-NC-SA 4.0.

Conflicts of interest

The authors declare no known conflicts of interest.

References

Afzal, S.N., Haque, I., Ahmedani, M.S., Munir, M., Firdous, S.S., Rauf, A., Ahmed, I., Rattus, A.R. and Fayyaz, M., (2009) Resistance potential of wheat germplasm (*Triticum aestivum* L.) against stripe rust disease under rained climate of Pakistan. *Pakistan Journal of Botany*, 413, pp.1463–1475.

Ali, S. and Hodson, D., (2017) Wheat rust surveillance: Field disease scoring and sample collection for phenotyping and molecular genotyping. In: S. Periyannan, ed., *Wheat Rust Diseases*, Methods in Molecular Biology. [online] New York, NY: Springer New York, pp.3–11. Available at: http://link.springer.com/10.1007/978-1-4939-7249-4_1.

Ali, S., Leconte, M., Rahman, H., Saqib, M.S., Gladieux, P., Enjalbert, J. and de Vallavieille-Pope, C., (2014) A high virulence and pathotype diversity of *Puccinia striiformis* f.sp. *tritici* at its centre of diversity, the Himalayan region of Pakistan. *European Journal of Plant Pathology*, 1402, pp.275–290.

Ali, S., Shah, S.J.A., Rahman, H., Saqib, M.S., Ibrahim, M. and Sajjad, M., (2009) Variability in wheat yield under yellow rust pressure in Pakistan. *Turkish Journal of Agriculture and Forestry*, 33, pp.537–546.

Allen-Sader, C., Thurston, W., Meyer, M., Nure, E., Bacha, N., Alemayehu, Y., Stutt, R.O.J.H., Safka, D., Craig, A.P., Derso, E., Burgin, L.E., Millington, S.C., Hort, M.C., Hodson, D.P. and Gilligan, C.A., (2019) An early warning system to predict and mitigate wheat rust diseases in Ethiopia. *Environmental Research Letters*, 1411, p.115004.

Alomar, O., Batlle, A., Brunetti, J.M., García, R., Gil, R., Granollers, T., Jiménez, S., Laviña, A., Reverté, C., Riudavets, J. and Virgili-Gomà, J., (2016) Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats. *EFSA Supporting Publications*, [online] 1312. Available at: https://data.europa.eu/doi/10.2903/sp.efsa.2016.EN-1118.

ARRCC, (2022) *Asia Regional Resilience to a Changing Climate - Phase II - Climate services to avoid food security threatening crop disease epidemics in South Asia - final report*.

Baidya, S., Magar, P.B., Basnet, R., Aryal, L., Acharya, B., Gupta, G.R., Pant, K.R., Subedi, M., Karki, A., Joshi, P. and Bhandari, D., (2022) Pathotypes Diversity of *Puccinia striiformis* causing Yellow Rust of Wheat in Nepal. *Journal of the Plant Protection Society*, 701, pp.94–103.

Bentley, A.R., Donovan, J., Sonder, K., Baudron, F., Lewis, J.M., Voss, R., Rutsaert, P., Poole, N., Kamoun, S., Saunders, D.G.O., Hodson, D., Hughes, D.P., Negra, C., Ibba, M.I., Snapp, S., Sida, T.S., Jaleta, M., Tesfaye, K., Becker-Reshef, I. and Govaerts, B., (2022) Near- to long-term measures to stabilize global wheat supplies and food security. *Nature Food*, 37, pp.483–486.

Bhavani, S., Singh, R.P., Hodson, D.P., Huerta-Espino, J. and Randhawa, M.S., (2022) Wheat Rusts: Current Status, Prospects of Genetic Control and Integrated Approaches to Enhance Resistance Durability. In: M.P. Reynolds and H.-J. Braun, eds., *Wheat Improvement*. [online] Cham: Springer International Publishing, pp.125–141. Available at: https://link.springer.com/10.1007/978-3-030-90673-3_8.

Borlaug Global Rust Initiative (BGRI), (2020) Stripe rust hits wheat crop in Nepal. *BGRI - Borlaug Global Rust Initiative*. Available at: https://bgri.cornell.edu/stripe-rust-hits-wheat-crop-in-nepal/ [Accessed 24 Apr. 2024].

Brown, J.K.M. and Hovmøller, M.S., (2002) Aerial Dispersal of Pathogens on the Global and Continental Scales and Its Impact on Plant Disease. *Science*, 2975581, pp.537–541.

Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K.A., Giovani, B., Saunders, D.G.O., Kamoun, S., Legg, J.P., Verdier, V., Lessel, J., Neher, R.A., Day, R., Pardey, P., Gullino, M.L., Records, A.R., Bextine, B., Leach, J.E., Staiger, S. and Tohme, J., (2019) A global surveillance system for crop diseases. *Science*, 3646447, pp.1237–1239.

Chen, X., (2020) Pathogens which threaten food security: *Puccinia striiformis*, the wheat stripe rust pathogen. *Food Security*, 122, pp.239–251.

Daume, S., (2016) Mining Twitter to monitor invasive alien species — An analytical framework and sample information topologies. *Ecological Informatics*, 31, pp.70–82.

Diouf, R., Sarr, E.N., Sall, O., Birregah, B., Bousso, M. and Mbaye, S.N., (2019) Web scraping: State-of-the-art and areas of application. In: *2019 IEEE International Conference on Big Data (Big Data)*. [online] 2019 IEEE International Conference on Big Data (Big Data). Los Angeles, CA, USA: IEEE, pp.6040–6042. Available at: https://ieeexplore.ieee.org/document/9005594/.

Faisal, A.A., (2023a) *ARRCC scraped media dashboard*. [online] Available at: https://arrcc-viewer.onrender.com/ [Accessed 24 Apr. 2024].

Faisal, A.A., (2023b) *AsifAlFaisal/project-arrcc: Published version of the ARRCC dashboard*. Available at: https://zenodo.org/record/8024493.

FAO, IFAD, UNICEF, WFP and WHO, (2022) *The state of food security and nutrition in the world 2022. Repurposing food and agricultural policies to make healthy diets more affordable*. [online] Rome: FAO. Available at: http://www.fao.org/documents/card/en/c/cc0639en.

FAOSTAT, (2023) *FAOSTAT food and agriculture data*. Available at: https://www.fao.org/faostat/en/#home [Accessed 24 Apr. 2024].

Ghermandi, A. and Sinclair, M., (2019) Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55, pp.36–47.

Global Rust Reference Centre, (2024) *Successful surveillance results in early first detection of Ug99 in South Asia*. [online] Global Rust Reference Centre. Available at: https://agro.au.dk/forskning/internationale-platforme/wheatrust/news-and-events/news-item/artikel/successful-surveillance-results-in-early-first-detection-of-ug99-in-south-asia [Accessed 30 Apr. 2024].

Hovmøller, M.S., Patpour, M., Rodriguez-Algaba, J., Thach, T., Søren, C.K., Justesen, A.F. and Hansen, J.G., (2023) GRRC 2022 report of stem and yellow rust genotyping and race analyses. *Global Rust Reference Centre*. [online] Available at: https://agro.au.dk/forskning/internationale-platforme/wheatrust/news-and-events/news-item/artikel/grrc-report-of-yellow-and-stem-rust-genotyping-and-race-analyses-2022 [Accessed 24 Apr. 2024].

Indian Council on Agricultural Research (ICAR) - Indian Institute of Wheat and Barley Research (IIWBR), (2020a) *Mehtaensis July 2020*. [online] ICAR - Indian Institute of Wheat and Barley Research. Available at: https://iiwbr.org.in/wp-content/uploads/2023/09/Mehtaensis-40-2-July-2020.pdf [Accessed 24 Apr. 2024].

Indian Council on Agricultural Research (ICAR) - Indian Institute of Wheat and Barley Research (IIWBR), (2020b) *Wheat Crop Health Newsletter – December 2019*. [online] Available at: https://iiwbr.icar.gov.in/wp-content/uploads/2020/01/WCH-Newsletter-Dec.-2019.pdf [Accessed 21 Aug. 2023].

Indian Council on Agricultural Research (ICAR) - Indian Institute of Wheat and Barley Research (IIWBR), (2021) *Mehtaensis January 2021*. [online] ICAR - Indian Institute of Wheat and Barley Research. Available at: https://iiwbr.org.in/wp-content/uploads/2023/09/Mehtaensis-411-January-2021.pdf [Accessed 24 Apr. 2024].

International Food Policy Research Institute (IFPRI) and International Institute For Applied Systems Analysis (IIASA), (2015) *Global spatially-disaggregated crop production statistics data for 2005 version* 3.2. Available at: https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/DHXBJX.

Jahanbin, K., Rahmanian, F., Rahmanian, V. and Jahromi, A.S., (2019) Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health. *GMS Hygiene and Infection Control*; 14:Doc19. [online] Available at: https://www.egms.de/en/journals/dgkh/2019-14/dgkh000334.shtml [Accessed 29 Apr. 2024].

Jansen, K. and De La Cruz Bekema, J., (2023) The control of transboundary plant diseases and the problem of the public good: Lessons from Fusarium wilt in banana. *NJAS: Impact in Agricultural and Life Sciences*, 951, p.2261402.

Jin, Y., Szabo, L.J. and Carson, M., (2010) Century-Old Mystery of *Puccinia striiformis* Life History Solved with the Identification of *Berberis* as an Alternate Host. *Phytopathology*, 1005, pp.432–435.

Jones, A., Thomson, D., Hort, M. and Devenish, B., (2007) The UK Met Office's next-generation atmospheric dispersion model, NAME III. In: *Air Pollution Modeling and Its Application* XVII. [online] Springer US, pp.580–589. Available at: http://link.springer.com/chapter/10.1007/978-0-387-68854-1_62.

Khan, M.R., Rehman, Z., Nazir, S.N., Tshewang, S., Baidya, S., Hodson, D., Imtiaz, M. and Ali, S., (2019) Genetic Divergence and Diversity in Himalayan *Puccinia striiformis* Populations from Bhutan, Nepal, and Pakistan. *Phytopathology*, 10910, pp.1793–1800.

Kisana, Mujahid, Y.M. and Mustafa, Z.S., (2003) *Wheat production and productivity* 2002- 2003: A *technical report to appraise the issues and future strategies*. Islamabad: National Agricultural Research Center, Pakistan Agricultural Research Council, Islamabad., p.19.

Mehmood, S., Sajid, M., Zhao, J., Huang, L. and Kang, Z., (2020) Alternate Hosts of *Puccinia striiformis* f. sp. *tritici* and Their Role. *Pathogens*, 96, p.434.

Met Office, (2013) *Operational numerical weather prediction* (NWP) *output from the global atmospheric part of the Met Office Unified Model* (UM). Available at: http://catalogue.ceda.ac.uk/uuid/41f061e11217e549a498971725e90520 [Accessed 24 Apr. 2024].

Meyer, M., Burgin, L., Hort, M.C., Hodson, D.P. and Gilligan, C.A., (2017a) Large-Scale Atmospheric Dispersal Simulations Identify Likely Airborne Incursion Routes of Wheat Stem Rust Into Ethiopia. *Phytopathology*, 10710, pp.1175–1186.

Meyer, M., Cox, J.A., Hitchings, M.D.T., Burgin, L., Hort, M.C., Hodson, D.P. and Gilligan, C.A., (2017b) Quantifying airborne dispersal routes of pathogens over continents to safeguard global wheat supply. *Nature Plants*, 310, pp.780–786.

Mitchell, R., (2015) *Web scraping with Python: collecting data from the modern web*. First Edition ed. Sebastopol, CA: O'Reilly Media.

Morris, C.E., Géniaux, G., Nédellec, C., Sauvion, N. and Soubeyrand, S., (2022) One Health concepts and challenges for surveillance, forecasting, and mitigation of plant disease beyond the traditional scope of crop production. *Plant Pathology*, 711, pp.86–97.

ODK, (2023) *OpenDataKit*. [online] Available at: https://getodk.org/ [Accessed 24 Apr. 2024].

Pilipiec, P., Samsten, I. and Bota, A., (2023) Surveillance of communicable diseases using social media: A systematic review. *PLOS ONE*, 182, p.e0282101.

Radici, A., Martinetti, D. and Bevacqua, D., (2022) Early-detection surveillance for stem rust of wheat: Insights from a global epidemic network based on airborne connectivity and host phenology. *Environmental Research Letters*, 176, p.064045.

Radici, A., Martinetti, D. and Bevacqua, D., (2023) Global benefits and domestic costs of a cooperative surveillance strategy to control transboundary crop pathogens. *PLANTS, PEOPLE, PLANET*, p.ppp3.10379.

RustTracker, (2023) *RustTracker.org - A Global Wheat Rust Monitoring System*. [online] RustTracker.org - A Global Wheat Rust Monitoring System. Available at: https://rusttracker.cimmyt.org/ [Accessed 24 Apr. 2024].

Saari, E.E. and Prescott, J.M., (1985) World Distribution in Relation to Economic Losses. In: *Diseases, Distribution, Epidemiology, and Control*. [online] Elsevier, pp.259–298. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780121484026500171.

Singh, R.P., Hodson, D.P., Jin, Y., Lagudah, E.S., Ayliffe, M.A., Bhavani, S., Rouse, M.N., Pretorius, Z.A., Szabo, L.J., Huerta-Espino, J., Basnet, B.R., Lan, C. and Hovmøller, M.S., (2015) Emergence and Spread of New Races of Wheat Stem Rust Fungus: Continued Threat to Food Security and Prospects of Genetic Control. *Phytopathology®*, 1057, pp.872–884.

Smith, J.W., (2023) *Advancing crop disease early warning in South Asia by complementing expert surveys with internet media scraping - supporting data and analysis*. Available at: https://zenodo.org/record/8296746 [Accessed 24 Apr. 2024].

TAFSSA, (2023) *Transforming Agrifood Systems in South Asia*. [online] Available at: https://www.cgiar.org/initiative/transforming-agrifood-systems-in-south-asia-tafssa/ [Accessed 24 Apr. 2024].

Thompson, R.N., Cobb, R.C., Gilligan, C.A. and Cunniffe, N.J., (2016) Management of invading pathogens should be informed by epidemiology rather than administrative boundaries. *Ecological Modelling*, 324, pp.28–32.

Walters, D., Baran, A.J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Manners, J., Morcrette, C., Mulcahy, J., Sanchez, C., Smith, C., Stratton, R., Tennant, W., Tomassini, L., Van Weverberg, K., Vosper, S., Willett, M., Browse, J., Bushell, A., Carslaw, K., Dalvi, M., Essery, R., Gedney, N., Hardiman, S., Johnson, B., Johnson, C., Jones, A., Jones, C., Mann, G., Milton, S., Rumbold, H., Sellar, A., Ujiie, M., Whitall, M., Williams, K. and Zerroukat, M., (2019) The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations. *Geoscientific Model Development*, 125, pp.1909–1963.

World Bank, (2021) *Agriculture, forestry, and fishing, value added (% of GDP)*. [World Development Indicators] The World Bank Group. Available at: https://databank.worldbank.org/source/world-development-indicators/preview/on [Accessed 24 Apr. 2024].