KAIQUE DOS SANTOS ALVES

DISENTANGLING ENVIRONMENTAL EFFECTS ON PLANT DISEASE EPIDEMICS AT THE REGIONAL SCALE

Tese apresentada à Universidade Federal de Viçosa como parte das exigências do Programa de Pós-Graduação em Fitopatologia, para obtenção do título de *Doctor Scientiae*.

Orientador: Emerson Medeiros Del Ponte

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
A474d
2022

Alves, Kaique dos Santos, 1995-
Disentangling environmental effects on plant disease epidemics at the regional scale / Kaique dos Santos Alves. – Viçosa, MG, 2022.
1 tese eletrônica (84 f.): il. (algumas color.).

Orientador: Emerson Medeiros Del Ponte.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Fitopatologia, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.053
Modo de acesso: World Wide Web.

1. Plantas - Doenças e pragas - Fatores climáticos. 2. Huanglongbing. 3. Frutas cítricas - Doenças e pragas. 4. Ferrugem da soja (Doença). 5. Mofos brancos. 6. Vagem - Doenças e pragas. I. Del Ponte, Emerson Medeiros, 1973-. II. Universidade Federal de Viçosa. Departamento de Fitopatologia. Programa de Pós-Graduação em Fitopatologia. III. Título.

CDD 22. ed. 571.92

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

KAIQUE DOS SANTOS ALVES

DISENTANGLING ENVIRONMENTAL EFFECTS ON PLANT DISEASE EPIDEMICS AT THE REGIONAL SCALE

Tese apresentada à Universidade Federal de Viçosa como parte das exigências do Programa de Pós-Graduação em Fitopatologia, para obtenção do título de *Doctor Scientiae*.

APROVADA: 17 de novembro de 2022.

Assentimento:

gov.br Documento assinado digitalmente
KAIQUE DOS SANTOS ALVES
Data: 16/02/2023 14:54:41-0300
Verifique em https://verificador.iti.b

Kaique dos Santos Alves
Autor

gov.br Documento assinado digitalmente
EMERSON MEDEIROS DEL PONTE
Data: 16/02/2023 15:46:19-0300
Verifique em https://verificador.iti.b

Emerson Medeiros Del Ponte
Orientador

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my parents, Cidileia dos Santos Alves e João Alves Neto;

I am grateful to my Advisor, Prof. Emerson Medeiros Del Ponte, for his exceptional mentoring, trust, and friendship during my master and doctorate degree.

I am also grateful to Dr. Sarah Jane Pethybridge for her outstanding mentoring during my internship in her research program at Cornell Agritech and also for providing funding for my internship;

Many thanks to Dr. Denis Antony Shah for his mentoring and collaboration on the research conducted in this project;

I would like to thank the Thesis examination committee members for their time and criticism;

I thank Universidade Federal de Viçosa, the Departamento de Fitopatologia and the Programa de Pós-graduação em Fitopatologia;

I am grateful to Cornell University and the Plant Pathology & Plant-Microbe Biology Section for hosting me during my research internship at Cornell Agritech;

I thank the current and past members of the plant disease epidemiology laboratory led by Prof. Emerson Medeiros Del Ponte for their friendship;

I thank the member of the *Epidemiology of VegetAble DisEases* (EVADE) *laboratory* for welcoming me during my internship;

I am grateful to the friends I made at Universidade Federal de Viçosa and at Cornell Agritech;

I am thankful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for providing the scholarship for my doctoral degree (Código de Financiamento 001).

"Chemical industry and plant breeders have forged fine tactical weapons; but only epidemiology sets the strategy." **(J. E. Vanderplank)**

ABSTRACT

ALVES, Kaique dos Santos, D.Sc., Universidade Federal de Viçosa, November, 2022. **Disentangling environmental effects on plant disease epidemics at the regional scale**. Adviser: Emerson Medeiros Del Ponte.

The environment plays an essential role in driving the occurrence and dynamics of plant disease epidemics. The weather is known to influence pathogen and host biology and should modulate the stages of the disease cycle. On the other hand, climate, the average of weather within long periods of time (i.e. 30 years), should set the predominance of pathogens and host genotypes across regions, and therefore the spatial distribution of plant diseases and their respective intensities. In this thesis, three studies aiming to unravel the effects of environmental factors on plant disease epidemics will be presented. The first study will demonstrate how the climate shapes the spatial distribution of citrus Huanglongbing prevalence in Minas Gerais. In the second study, the time to onset of soybean rust in commercial soybean fields from Southern Brazil was associated with the El Nino Southern Oscillation, a phenomenon that triggers extreme weather events around the globe and also in Brazil; The third study integrates cutting-edge statistical methodology to associate weather time series and soil properties data to white mold prevalence in snap bean fields in New York, United States. The results led to novel insights into pathogen biology and disease risk at the regional and local scales for the three pathosystems under study.

Keywords: Epidemiology. Huanglongbing. Soybean rust. White mold. Bayesian. Machine learning.

RESUMO

ALVES, Kaique dos Santos, D.Sc., Universidade Federal de Viçosa, novembro de 2022. **Desembaraçando efeitos ambientais em epidemias de doenças de plantas em escalas regionais**. Orientador: Emerson Medeiros Del Ponte.

O ambiente desempenha um papel essencial para ocorrência e dinâmica de epidemias de doenças de plantas. O tempo (meteorológico) é conhecido por influenciar a biologia do patógeno e do hospedeiro e por isso desempenha papel na modulação dos estágios do ciclo das doenças. Por outro lado, o clima, dado como a média do tempo (meteorológico) durante longos períodos (30 anos), deve definir a predominância de patógenos e genótipos hospedeiros nas regiões e, portanto, a distribuição espacial das doenças de plantas e suas respectivas intensidades. Nesta tese, serão apresentados três estudos com o objetivo de desvendar os efeitos de fatores ambientais em epidemias de doenças de plantas. O primeiro estudo demonstrará como o clima molda a distribuição espacial da prevalência do Huanglongbing (também conhecido como *greening*) dos citros em Minas Gerais. No segundo estudo, o tempo de aparecimento da ferrugem da soja em lavouras comerciais de soja do Sul do Brasil foi associado ao El Nino-Southern Oscillation, fenômeno que desencadeia eventos climáticos extremos ao redor do globo e também no Brasil; O terceiro estudo integra metodologia estatística de ponta para associar séries temporais de variáveis meteorológicas e dados de propriedades do solo à prevalência de mofo branco em campos de feijão-vagem em Nova York, Estados Unidos. Os resultados levaram a novos entendimentos sobre a biologia dos patógenos e o risco de doenças nas escalas regional e local para os três patossistemas em estudo.

Palavras-chave: Epidemiologia. Huanglongbing. Ferrugem da soja. Mofo branco. Bayesiano. Aprendizado de máquina.

TABLE OF CONTENTS

GENERAL INTRODUCTION

Environmental factors are key components driving plant disease epidemics. This is because environmental factors have direct and indirect effects on both host plant and pathogen populations, a conceptual framework classically represented by the disease triangle (Scholthof 2007). Among the main components of the environment are the climate, weather, and soil properties. Therefore, significant effort has been made to understand and predict disease outbreaks from weather, or climate-related variables. Contrasting the differences between weather and climate will help better understand how the environment affects plant diseases and their distribution in the agricultural landscape. Weather refers to the state of the atmosphere in a short period of time. Climate refers to the average of the weather over time (~30 years) and space (Gutro 2015; NOAA 2021). Therefore, weather can drive epidemic dynamics prior to or during cropping seasons by regulating the stages of the infection cycle, plant growth, or management efficacy. On the other hand, climate can drive the spatial distribution of pathogens and their diseases, and the dominance of pathogen genotypes in crops across large (regional or global) scales. Climatic conditions at a local scale are usually well characterized (at some level). Therefore, most of the uncertainty in disease risk at the field scale can be considered a result of the volatility in local weather. Over many years, the study of the effects of weather on plant diseases at a field scale has revealed the importance of key variables such as temperature, relative humidity, leaf wetness, and rainfall for forecasting disease risk (Bourke 1970; Del Ponte and Esker 2008; Gent et al. 2013).

Disease-related variables and their resultant effects on crop yield can also be affected by fluctuations in climate patterns in a geographical region. Any anomaly in the expected (normal) climate might lead to an unpredictable disruption in the pathosystem status quo. Fluctuations in climate in one region can be the result of climate variations in another region of the planet. This phenomenon is called teleconnection (Kriss et al. 2012; Troccoli 2010). An example of teleconnection is the El Niño–Southern Oscillation (ENSO), which is responsible for triggering extreme climatic events around the globe. For instance, El Niño events can increase or decrease rainfall levels in southern Brazil (subtropics) and droughts in the tropics during the spring and summer months (Grimm 2003; Cirino et al. 2015). Moreover, the rise in global temperature due to the increase in atmospheric CO_2 concentration is likely to intensify climate variability and increase the number of extreme weather-related events (Seneviratne et

al. 2012). Hence, studying the impact of extreme events and climate change on plant disease epidemics is vital (Bebber 2019; Chaloner et al. 2021; Fones et al. 2020).

This doctorate thesis has three research chapters, each one tackling a different aspect of the environment: Climate, weather, and teleconnections. The first chapter is an already-published article that demonstrates the spatial distribution of Huanglongbing, a severe citrus disease, is governed by climatic factors, more specifically winter temperatures and windy speed. In the second chapter, the time for soybean rust outbreaks in commercial soybean fields from southern Brazil is associated with temperature anomalies in the Pacific Ocean within the Niño 3.4 region, which is used for classifying the ENSO phases. Finally, in the third chapter, a new framework for studying weather time series and soil properties variables effects on plant diseases is developed and applied to investigate how such environmental factors influence white mold prevalence in Snap bean fields in New York, United States. The results of these research works provide new insights into plant disease management and monitoring, as well as providing new resources and tools that could be widely used in epidemiological studies.

REFERENCES

Bebber, D. P. 2019. Climate change effects on Black Sigatoka disease of banana. Philosophical Transactions of the Royal Society B: Biological Sciences. 374:20180269.

Bourke, P. M. A. 1970. Use of weather information in the prediction of plant disease epiphytotics. Annual Review of Phytopathology. 8:345–370.

Chaloner, T. M., Gurr, S. J., and Bebber, D. P. 2021. Plant pathogen infection risk tracks global crop yields under climate change. Nat. Clim. Chang. 11:710–715.

Cirino, P. H., Féres, J. G., Braga, M. J., and Reis, E. 2015. Assessing the Impacts of ENSO-related Weather Effects on the Brazilian Agriculture. Procedia Economics and Finance. 24:146–155.

Del Ponte, E. M., and Esker, P. D. 2008. Meteorological factors and Asian soybean rust epidemics: a systems approach and implications for risk assessment. Scientia Agricola. 65:88–97.

Fones, H. N., Bebber, D. P., Chaloner, T. M., Kay, W. T., Steinberg, G., and Gurr, S. J. 2020. Threats to global food security from emerging fungal and oomycete crop pathogens. Nature Food. 1:332–342.

Gent, D. H., Mahaffee, W. F., McRoberts, N., and Pfender, W. F. 2013. The use and role of predictive systems in disease management. Annual Review of Phytopathology. 51:267–289.

Grimm, A. M. 2003. The El Niño Impact on the Summer Monsoon in Brazil: Regional Processes versus Remote Influences. Journal of Climate. 16:263–280.

Gutro, R. 2015. What's the Difference Between Weather and Climate? National Aeronautics and Space Administration (NASA). Available at: http://www.nasa.gov/mission_pages/noaa-n/climate/climate_weather.html [Accessed February 23, 2022].

Kriss, A. B., Paul, P. A., and Madden, L. V. 2012. Variability in Fusarium Head Blight Epidemics in Relation to Global Climate Fluctuations as Represented by the El Niño-Southern Oscillation and Other Atmospheric Patterns. Phytopathology®. 102:55–64.

NOAA. 2021. What is the difference between weather and climate? National Ocean Service website. Available at: https://oceanservice.noaa.gov/facts/weather_climate.html [Accessed February 23, 2022].

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., et al. 2012. Changes in Climate Extremes and their Impacts on the Natural Physical Environment. In *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, eds. Christopher B. Field, Vicente Barros, Thomas F. Stocker, and Qin Dahe. Cambridge: Cambridge University Press, p. 109–230. Available at: https://www.cambridge.org/core/product/identifier/CBO9781139177245A030/type/book_part [Accessed March 15, 2022].

Scholthof, K.-B. G. 2007. The disease triangle: pathogens, the environment and society. Nat Rev Microbiol. 5:152–156.

Troccoli, A. 2010. Seasonal climate forecasting. Meteorological Applications. 17:251–268.

CHAPTER 1: Linking Climate Variables to Large-Scale Spatial Pattern and Risk of Citrus Huanglongbing: A Hierarchical Bayesian Modeling Approach

This chapter is published: Alves, K. S., Rothmann, L. A., and Del Ponte, E. M. 2022. Linking Climate Variables to Large-Scale Spatial Pattern and Risk of Citrus Huanglongbing: A Hierarchical Bayesian Modeling Approach. Phytopathology®. 112:189–196.

ABSTRACT

Huanglongbing (HLB) is one of the most important diseases affecting citriculture in the world. Knowledge of climatic factors linked to HLB risk at large spatial scales is limited. We gathered HLB presence/absence data from official surveys conducted in the state of Minas Gerais, Brazil, over 13 years. The total count of orange and mandarin orchards, and mean orchard area, normalized to a spatial grid of 60 cells (55 × 55 km), were derived from the same database. Monthly climate normals (1984 to 2013) of rainfall, mean temperature, and wind speed split into *rainy* (September to April) and *dry* (May to August) seasons (*annual* summary was retained) were obtained for each grid cell. Two hierarchical Bayesian modeling approaches were evaluated both based on the integrated nested Laplace approximation methodology. The first, the climate covariates model (CC model), used orchard, climate, and the spatial effect as covariates. The second, principal components (PC model), used the first three components from a principal component analysis of all variables and the spatial effect as covariates. Both models showed an inverse relationship between posterior prevalence and grid cell mean temperature during the dry season. Annual wind speed, as well as annual and rainy season rainfall, contributed to HLB risk, in the CC and PC models, respectively. A partial influence of neighboring regions on HLB risk was observed. The results should assist policymakers in defining regions at HLB risk and guide monitoring strategies to mitigate further spread of HLB in the state of Minas Gerais.

Keywords: Citrus greening, epidemiology, disease risk, mapping, sweet orange, epidemic

INTRODUCTION

Huanglongbing (HLB), also referred to as citrus greening, is among the most destructive of citrus diseases worldwide, including in Brazil (Gottwald et al. 2010; Bassanezi et al. 2020). In Brazil, HLB was first reported in São Paulo (SP) state during the 2004 season (Coletta-Filho et al. 2004; do Carmo Teixeira et al. 2005). Thirteen years later, the incidence of HLB in sweet orange orchards in the citrus belt (spanning SP and Minas Gerais; MG states) had increased to ~19%, with an estimated 37 million trees infected (Fundecitrus 2019). In 2018, over 57% of orchards, including mandarins and sweet oranges in MG were affected by HLB (Alves et al. 2020). A recent 10-year analysis showed that the epidemic front of HLB was advancing at 25.7 km per year in central and south MG, and advancing at as much as 45.9 km per year in the border region of SP, respectively (Alves et al. 2020).

Three Gram-negative bacterium species of the genus *Candidatus* Liberibacter can cause HLB; namely, *Ca*. L. africanus (Jagoueix et al. 1994), *Ca*. L. americanus (Clam) (Coletta-Filho et al. 2004; do Carmo Teixeira et al. 2005) and *Ca*. L. asiaticus (CLas) (Jagoueix et al. 1994). The predominant species in Brazil was *Ca*. L. americanus, until 2008 when a shift in the population occurred and CLas became the most prevalent species (>99.9%) found in orchards (Lopes et al. 2009; Bassanezi et al. 2020). The bacteria exhibit persistent propagative transmission with the phloem-probing psyllid vectors, both with the nymphs, and adults, which are mainly responsible for spread and transmission of the bacteria under natural conditions (Ammar et al. 2016).

The Asian citrus psyllid, *Diaphorina citri*, is the primary vector responsible for the spread and transmission of CLam and CLas (do Carmo Teixeira et al., 2005; Bové, 2006). The whole psyllid life-cycle, including egg, nymph, and adult, encompassing pathogen acquisition, replication, transmission, survival, and spread, should be considered as factors affecting the onset and spread of HLB (Razi et al. 2014; Teck et al. 2011). Development of epidemic intervention strategies, and their successful implementation is therefore reliant on determining factors driving disease development in the plant, and spread between plants (Madden et al. 2007).

The prevalence of HLB in different states in Brazil can be attributed to the psyllid vectors' capability for short- (plant-to-plant and within orchard; 5-320 m) and long-distance (between orchard; ~2.5 km) movement via passive and active means, i.e., wind-assisted dispersal and consecutive short flights, respectively (Carmo-Sousa et al. 2020; Antolinez et al. 2021). Additionally, the use of infected propagative material for grafting has also been

demonstrated to transmit HLB and can thus contribute to the dissemination of the pathogen (van Vuuren & da Graça 1993). Orchard-level studies indicate that clustering of HLB-infected trees occurs as a result of vector movement, with an edge effect due to trees becoming infected along an orchard border, which can subsequently lead to nearby trees becoming infected and enlarging the initial foci, or generating new foci within the orchard (Bassanezi et al. 2005). At a regional level, an aggregated spatial pattern of HLB was observed in orchards in MG, with a more pronounced aggregation in mandarin (*Citrus reticulata* Blanco) orchards compared to sweet orange (*C. sinensis* Osbeck) orchards (Alves et al. 2020).

Due to HLB's polyetic nature, quantitative epidemiological studies can be difficult to implement - multi-season data are required and eradication of the symptomatic plants reduces the opportunity to continuously monitor disease progress (Gottwald 2010). Where multi-year field studies are conducted, numerous environmental factors may be associated with the spatial and temporal progress of HLB. Rainfall was positively correlated with HLB incidence in Florida, indicating that higher rainfall is associated with a greater risk of HLB infection, likely due to a plant physiological response or effects on vector population dynamics (Shimwela et al. 2018). The psyllid vector of the HLB bacterium has a narrower temperature range tolerance than that of the bacterium itself (Gutierrez and Ponti 2013). Furthermore, within the species complex, CLas tolerates higher temperatures than *Ca*. L. americanus, which is assumed to be part of the reason for the shift in species dominance (Lopes et al. 2009; Gasparoto et al. 2012). The host as an environment also need to be considered. For example, newly emerging leaves (flushes) have been associated with higher densities of the psyllid vector in citrus orchards (Gutierrez and Ponti 2013; Lewis-Rosenblum et al. 2015).

The mitigation of HLB is limited to preventative measures and include eradication and exclusion strategies, quarantine, planting certified-healthy nursery trees, and insecticide applications to suppress vector populations and bacterial dissemination (Bové 2006; Bassanezi et al. 2020). Eradication of symptomatic trees aims to reduce inoculum reservoirs in order to mitigate losses and the spread of HLB, which comes at a high cost to growers (Lopes et al. 2008). In response to the HLB epidemic in Brazil, statewide initiatives were implemented in 2004 to monitor the occurrence of HLB and enforce orchard clearing when the incidence of HLB exceeded a 28% threshold (Craig et al. 2018; Bassanezi et al. 2020). Ultimately, a regional disease management approach is the most appropriate tool for effective management of HLB, which must include systematic and coordinated vector control and removal of inoculum sources (symptomatic citrus trees) from commercial and noncommercial properties (Bassanezi et al. 2013).

MG ranks third in citrus production in Brazil (IBGE 2018 cited in Carvalho et al. 2019) and is in one of the four major citrus-producing regions of the country (Passos et al. 2018). We hypothesize that climatic variables (temperature, rainfall, and wind speed) and orchard-related variables, such as orchard size and citrus species, in part explain the relative geographical prevalence of HLB in MG. In this study, we aimed to link climatic factors to the present spatial distribution of HLB prevalence in MG using hierarchical Bayesian spatial modeling. Bayesian spatial modeling should enable quantification and visualization of the disease spatial distribution, including the effect of a range of covariates and random spatial effects. The methodology is similar to a recent study that modeled the effect of climatic variables on the spatial distribution of *Xylella fastidiosa* in regions of Spain and Italy (Cendoya et al. 2020). The results could be used to assist policy makers, plant health extension specialists, and producers in monitoring HLB and, if needed, optimizing management strategies to specific areas based on the probability of HLB prevalence.

MATERIAL AND METHODS

HLB prevalence

The status of HLB (presence or absence) in citrus orchards has been regularly monitored by the Instituto Mineiro de Agropecuaria (IMA, Belo Horizonte, MG) in MG since the first detection of the disease in 2005. IMA surveys all citrus orchards across municipalities and categorizes risk for HLB. A risk 1 category is associated with municipalities where HLB has already been detected in any citrus orchard. Neighboring municipalities where HLB has not yet been detected are a risk 2 category. All citrus orchards located in the municipalities under risks 1 and 2 are monitored, if risk categories are not attributed to municipalities the orchards within those municipalities are not monitored. Each citrus plant in all orchards is inspected for HLB and HLB-like symptoms by trained inspectors. Once an HLB-symptomatic plant is detected, samples are sent to a laboratory for confirmation by polymerase chain reaction (PCR) assays. In this study, an orchard was categorized as “HLB-present” (positive) if HLB had been detected between 2005 and 2018, otherwise, the orchard was classified as “HLB-absent” (negative).

All positive and negative (HLB-present or absent) orchards are included in the database with their respective location (latitude and longitude) and other attributes related to the orchard (citrus species, orchard area, HLB cumulative incidence, number of trees eradicated, total number of trees). A full description of the data is available elsewhere (Alves et al. 2020) and is

available in a citable research compendium (osf.io/23ghm). Data were compiled into a spatial grid of 0.5 × 0.5 decimal degrees (the smallest size for all grid cells to have at least one neighboring cell), which is equivalent to a 55.5 × 55.5 km grid at the Equator (Fig. 1).

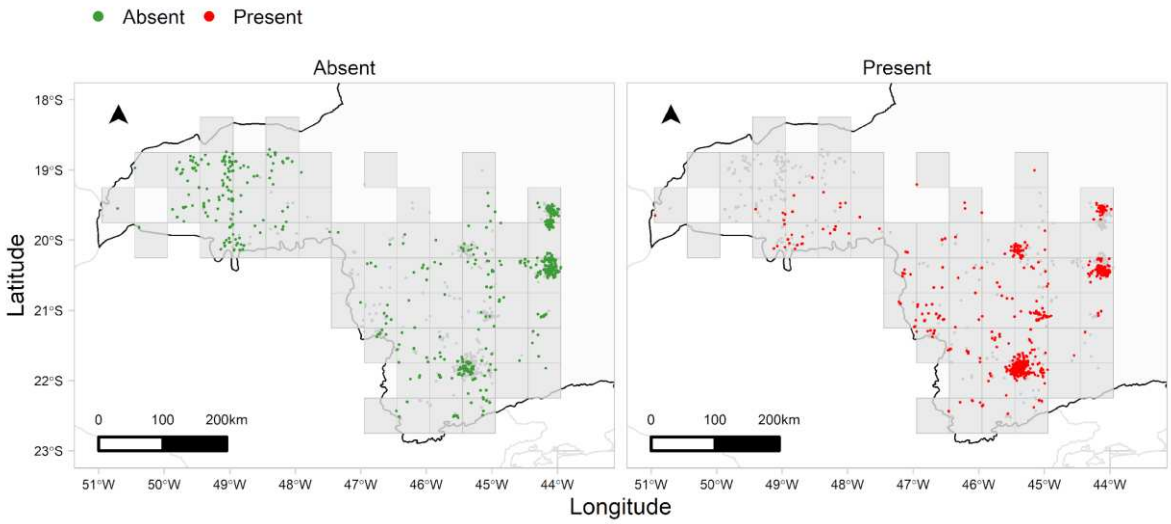

Fig. 1. Spatial distribution of citrus orchards with and without huanglongbing from 2005 to 2018 in Minas Gerais, Brazil. Gray squares represent the 55.5 × 55.5 km grid cells in which all data were based.

Previous reports suggest an effect of orchard area on temporal and spatial disease progress due to the orchard edge effect - from where the disease tends to spread into the orchard (Gottwald et al. 2010; Leal et al. 2010; Shen et al. 2013; Gasparoto et al. 2018; Alves et al. 2020). Consequently, the mean orchard areas were calculated by grid cells. Since there was a large range in orchard areas, with a few very large orchards and many smaller orchards, area was log-transformed prior to spatial modeling. The number of mandarin and orange orchards in each grid cell was calculated as recent work demonstrated more mandarins are cultivated in the south region of MG compared to the west of the state, where sweet orange orchards predominate (Alves et al. 2020). The orchard type data was log transformed due to the presence of high values in some grid cells causing heterogeneity of variance. To avoid issues of log-transforming zero count data, we added 1 to all values of orchard type frequency.

Climate data

Climate data were obtained using the R package (R Core Team, 2021) *nasapower* (Sparks 2018). The *nasapower* package provides capability to download data from the NASA POWER (National Aeronautics and Space Administration Prediction of Worldwide Energy Resource, Washington DC) database, which is freely available for download. Climatic data

(long-term monthly average and annual mean data for a 30-year period from Jan. 1984 to Dec. 2013 included temperature at two meters (T, °C), precipitation (rainfall) (*P*, mm/day) and wind speed at 10 m (W, m/s). The data obtained corresponded to the coordinates in the center of each 55.5 × 55.5 km grid cell. To account for seasonal variations in the variables during the year, mean values of climatic variables during the rainy and dry seasons were calculated to obtain mean temperatures (T_{rainy} and T_{dry}), mean rainfall (P_{rainy} and P_{dry}), and mean wind speed (W_{rainy} and W_{dry}) during the rainy and dry seasons in MG. The averages for the rainy season were calculated from September to April (the end of winter to the beginning of autumn; de Sá Júnior et al. 2012), while the averages for the dry season were calculated from May to August (the middle of autumn to the middle of winter; de Sá Júnior et al. 2012).

Predictor variables

The candidate predictor variables were the log-transformed mean orchard area (*A*), the log-transformed number of mandarin and orange orchards ($N_{mandarins}$ and N_{orange}), T_{rainy}, T_{dry}, P_{rainy}, P_{dry}, W_{rainy}, W_{dry}, mean annual wind speed (W_{annual}), temperature (T_{annual}), and rainfall (P_{annual}). Subsequently, only a subset of these variables was used in the modeling procedure, described below under "Model selection".

Additionally, a principal component analysis (PCA) of HLB prevalence in relation to the variables described above was performed. This method was used to combine multiple covariates into a reduced number of uncorrelated principal components (PCs) to use as predictor variables. To account for different metrics across different variables, the PCA was performed based on the correlation matrix, in which the correlation of each variable with the PCs was expressed by a rotation with the Varimax method (Cendoya et al. 2020). The first three PCs were used as candidate predictors. The function 'principal()' from the *psych* package (Revelle 2020) was used to perform the PCA.

Modeling framework

A Bayesian hierarchical approach, similar to one used to analyze the spatial distribution of *X. fastidiosa* in Spain and Italy (Cendoya et al. 2020), was applied to the HLB data. An integrated nested Laplace approximations (INLA) computation method was used to obtain posterior probability distributions of the model's parameter and hyperparameters. The analysis was performed using the *inla*() function of the R package INLA (Rue et al. 2009).

The spatial model used was a reparameterization of the Besag-York-Mollié model (BYM-model) (Besag et al. 1991), which is commonly used in disease mapping (Simpson et al. 2017). The objective is to model the observed HLB prevalence (incidence of HLB-affected orchards) y_i in a grid cell i, where $i = 1, \ldots, G$, being G the number of grid cells. In our study y_i is assumed to follow a binomial distribution i.e., $y_i \sim Binomial(n_i, \pi_i)$, where n_i is the number of orchards in a grid cell and π_i is the probability of an orchard being affected by HLB. The original BYM-model for the number of orchards in a grid cell is $\eta_i = \beta_0 + X_i\beta_m + u_i + v_i$, where β_0 is the overall intercept, β_m is the effect of the covariates (or predictors) ($m = 1, \ldots, N_m$, N_m is the number of covariates), u_i is a zero-mean Gaussian with a precision matrix representing the unstructured random effect, and v_i is the random effects spatial component that accounts for the similarities of the neighboring grid cells (Besag 1974; Besag et al. 1991; Simpson et al. 2017; Cendoya et al. 2020). The reparametrized model (Equation 1) adds v_i^*, which is the scaled spatially structured component (Simpson et al. 2017):

$$\eta_i = \beta_0 + X_i\beta_m + \frac{1}{\tau}\left(\sqrt{1-\phi}u_i + \sqrt{\phi}v_i^*\right), \quad (1)$$

where $1/\tau$ is the marginal precision contribution for u_i and v_i^*. The mixing parameter ϕ ($0 \leq \phi \leq 1$), is the fraction of the variance explained by the spatial structure, represented by u_i and v_i^*. Higher values of ϕ indicate high dependence on the spatial component. X_i represents the vector of covariates. For further details regarding the reparameterized BYM-model, the reader is referred to Simpson *et al.* (2017).

As described above, the spatial grid used to aggregate HLB spatial data were defined as 0.5×0.5 decimal degrees. In order to obtain all grid cells with at least one neighboring cell, we converted grid coordinates to Universal Transverse Mercator (UTM) coordinates. Two grid cells were considered to be neighbors if they were at a maximum distance range of 100 km. Coordinates were transformed to UTM using the *spTransform*() function in the R package *sp* (Bivand et al. 2013).

The full model and respective priors for model parameter and hyperparameters are:

$$\begin{gathered}
\eta_i = \beta_0 + X_i\beta_m + \frac{1}{\tau}\left(\sqrt{1-\phi u_i} + \sqrt{\phi v_i^*}\right) \\
y_i \sim Binomial(n_i, \pi_i), i = 1, \ldots, G \\
logit(\pi_i) = \beta_0 + X_i\beta_m + \frac{1}{\tau}\left(\sqrt{1-\phi u_i} + \sqrt{\phi v_i^*}\right) \\
P(\beta_0) \propto 1 \\
\beta_m \sim N(\mu = 0, \tau = 10^{-3}), m = 1, \ldots, N_\beta
\end{gathered} \quad (2)$$

$$\tau \sim PCprior(0.5/0.31, 0.01),$$
$$\phi \sim PCprior(0.5, 2/3)$$

Where $PCprior(\cdot)$ defines the penalized complexity priors for the hyperparameters (Besag 1974; Besag et al. 1991; Simpson et al. 2017; Cendoya et al. 2020).

Model Selection

To identify the best candidate models, a total of 2^{13} models would need to be tested (where 13 is the total number of covariates plus the spatial random effect) resulting in a total of 8,192 models, which would require a lot of time and computational resources. To reduce the number of covariates and avoid multicollinearity, Pearson's correlations were performed among all candidate predictors. If the absolute correlation coefficient $|r|$ value was greater than 0.7, one of the variables was removed from subsequent analysis. The remaining variables were used in the modeling procedure, in which 2^k models (where k is the number of predictor variables including the spatial effect) were assessed using the Watanabe Akaike information criterion (WAIC; Watanabe, 2010) and the logarithmic conditional predictive ordinate (LCPO; Pettit, 1990). Models with lower WAIC and LCPO values were selected.

Data and code availability

All modeling procedures, downloading of climate data, and graphical work were performed using R version 4.0.2 (R Core Team 2021), the code was fully annotated using R markdown (Xie et al. 2018) to enable reproducibility. The data tidying and exploratory analysis (including maps) presented in the study were conducted using procedures available in the *tidyverse* package (Wickham et al. 2019). Details for installation of the *INLA* package can be accessed at www.r-inla.org/download-install. All files (data and codes) used in the study were organized as a research compendium stored in the Open Science Framework and can be freely accessed for download at https://osf.io/nyvak.

RESULTS

Climate covariates (CC) model

The variables selected based on the $|r| \leq 0.7$ pairwise threshold were the log-transformed area (A), the log-transformed number of mandarin and orange orchards ($N_{mandarins}$ and N_{orange}), the mean temperature during the dry season (T_{dry}), the mean rainfall during the rainy season (P_{rainy}), and the annual wind speed (W_{annual}). Distribution of values for these

variables ranged from -0.51 to 7.24 (A), 0 to 5.69 ($N_{mandarins}$), 0 to 3.55 (N_{orange}), 4.75 mm/day to 5.81 mm/day (P_{rainy}), 15.26°C to 22.39°C (T_{dry}), and 1.72 m/s to 3.56 m/s (W_{annual}). The spatial distribution maps of the orchard-related and climatic variables are presented (Fig. 2).

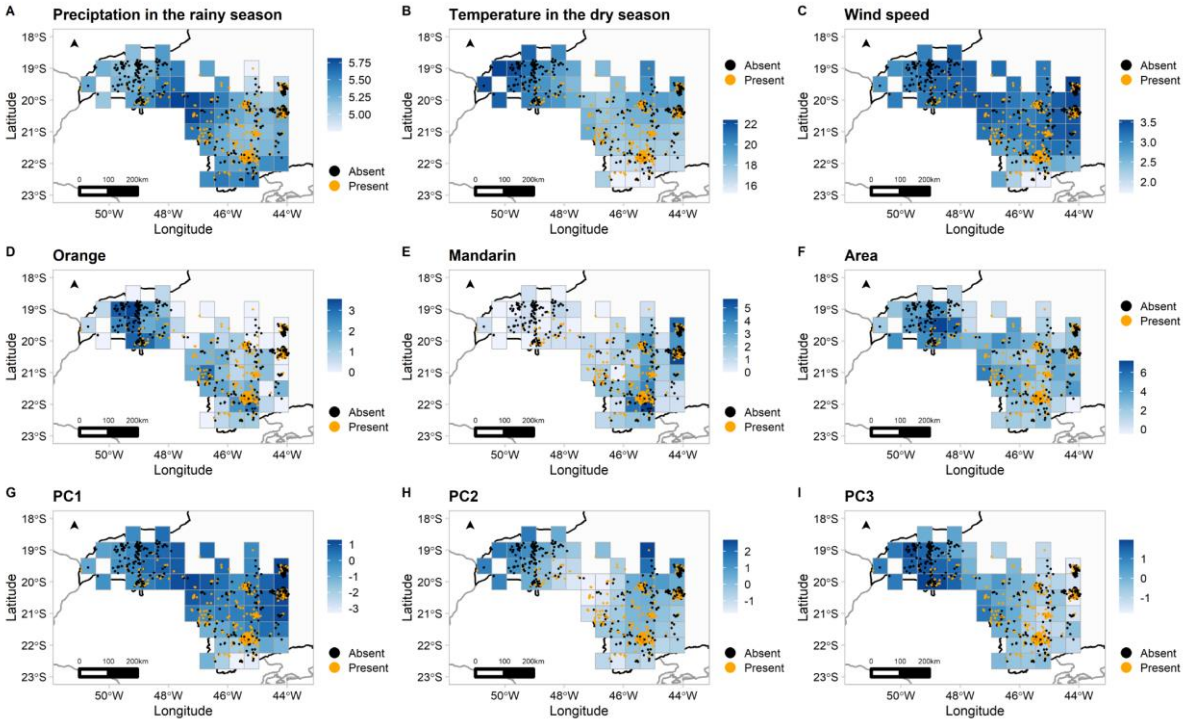

Fig. 2. Candidate variables used for modeling the spatial distribution of citrus huanglongbing in Minas Gerais. The candidate climate variables screened in the hierarchical Bayesian model were mean precipitation (rainfall) during the rainy season (A), mean temperature during the dry season (B), annual wind speed (C), log-transformed number of mandarin (D) and of orange orchards (E) log-transformed mean orchard area (F). The candidate variables tested in the hierarchical Bayesian models using principal components as covariates were the first three principal components (G, H, and I) that were based on all 12 original variables (see Table 2).

As six variables (A, $N_{mandarins}$, N_{orange}, T_{dry}, P_{rainy} and W_{annual}) and the spatial effect were selected, the number of models evaluated combining all variables was 2^7, or a total of 128 models. The model with lowest WAIC (215.678) and LCPO (1.421) was the model that combined P_{rainy}, T_{dry}, W_{annual}, and random spatial effect. Besides the mean of the posterior distribution of the effect parameter for P_{rainy} being positive, the 95% highest probability density (HPD, values ranging between HPD_{lower} and HPD_{upper}) interval contained zero, which provided insufficient evidence of a positive effect of P_{rainy} (Table 1). The mean of the posterior distribution for the effect of T_{dry} was negative, and the HPD interval did not contain zero, providing evidence of a lower probability of HLB in regions with high temperatures during the dry season, and *vice versa*. On the other hand, we found evidence of a positive effect of wind

speed on the distribution of HLB prevalence, since the posterior mean for the parameter of W_{annual} was positive and the HPD interval did not contain zero (Table 1). The mean of the posterior distribution for the mixing parameter ϕ was 0.141, indicating a partial influence of spatial effect on the probability of HLB presence. The spatial distribution of the posterior mean and standard deviation of the spatial effect are presented (Fig. 3A and 3B, respectively). The posterior mean of the spatial effect ranged from -2.68 to 2.06 across the grid, while the posterior standard deviation ranged from 0.36 to 1.68 across the grid. Higher posterior mean values for the spatial effect indicate a higher probability of HLB presence.

Table 1. Mean, median, standard deviation (SD), 95% highest posterior density (HPD) interval, and mode of model parameters and hyperparameters of the best model with climate covariates for citrus Huanglongbing prevalence distribution in Minas Gerais, Brazil.

Parameters[b]	**Mean**	**Median**	**SD**	**HPD$_{lower}$**	**HPD$_{upper}$**	**Mode**
β_0	1.244	1.245	8.309	-15.241	17.618	1.270
P_{rainy}	1.264	1.264	1.198	-1.102	3.625	1.263
[a] T_{dry}	-0.670	-0.666	0.211	-1.090	-0.257	-0.660
[a] W_{annual}	1.330	1.317	0.673	0.016	2.661	1.293
Hyperparameters	**Mean**	**Median**	**SD**	**HPD$_{lower}$**	**HPD$_{upper}$**	**Mode**
τ	0.678	0.652	0.194	0.339	1.064	0.603
ϕ	0.141	0.090	0.146	8.470×10^{-6}	0.451	0.010

[a] 95% HPD interval excludes 0.

[b] β_0 = model intercept; P_{rainy} = precipitation (rainfall) in the rainy season (September to April); T_{dry} = temperature in the dry season (May to August); W_{annual} = annual wind speed; τ = variance; ϕ = the mixing parameter (spatial effect).

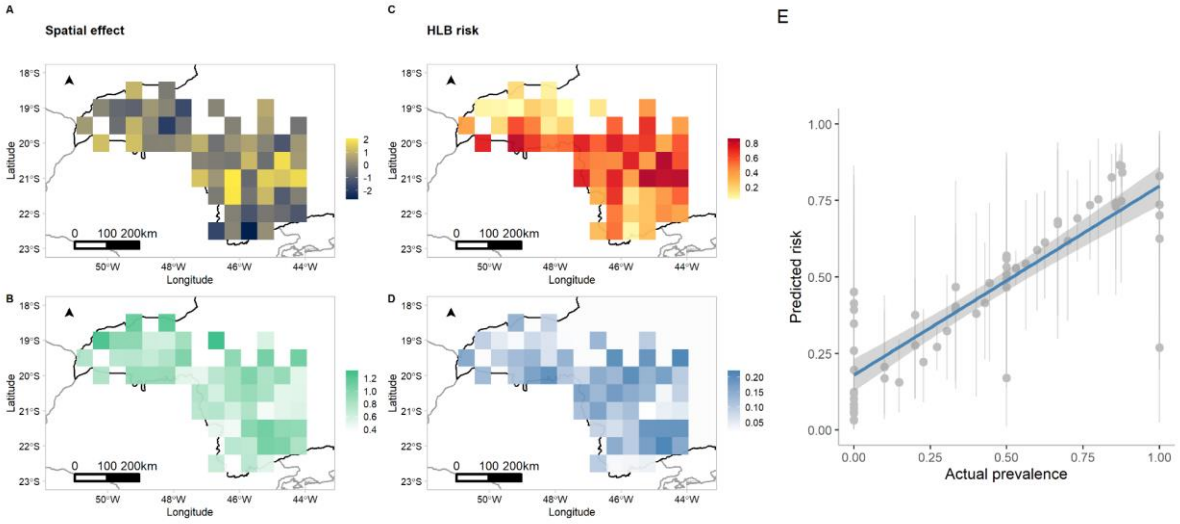

Fig. 3. Model with climate covariates and the spatial effect. Map (A) depicts the mean, and map (B) the standard deviation of the posterior distribution of the spatial effect. Map (C) depicts the posterior mean and map (D) the standard deviation of the spatial distribution of HLB prevalence (or risk) in Minas Gerais using the climate covariate model with mean daily precipitation (rainfall) during the rainy season (P_{rainy}), mean temperature during the dry season (T_{dry}), mean annual wind speed (W_{annual}), and the spatial effect as covariates. (E) shows the relationship between actual prevalence and the posterior mean (dots) and 95% highest posterior density (error bars) obtained with the climate covariate model.

The posterior distribution of HLB prevalence is given in terms of probability {0,1}, which may also be referred to as HLB risk. The spatial distribution of the posterior mean of HLB prevalence in MG based on the model with climate covariates and the spatial effect shows that the highest mean probability, 0.8650, was located in a region of high densities of citrus orchards in the eastern region (Fig. 3C and Fig. 1). In contrast, the lowest posterior mean probabilities, <0.05, were in peripheral grid cells to the northwest and south of the spatial grid, where most orchards were HLB-free (Fig. 3C). The posterior standard deviation of HLB prevalence ranged from 0.02 to 0.22 (Fig. 3D). Lower uncertainties were found on either low or high posterior means of HLB prevalence. Overall, there was a strong association between the estimated HLB risk and actual HLB prevalence (Fig. 3E; coefficient of determination (r^2 = 0.74)).

Principal components (PC) model

The first three principal components obtained in the PCA explained 81.1% of the total cumulative variation in the data (Table 2). The percentage of variance explained by the first, second, and third principal components were 30.7%, 28.4%, and 22.0%, respectively. Wind speed-related variables contributed the most to the first principal component (PC1), with coefficients greater than 0.88 (W_{rainy} contributed 0.974 to PC1). For the second principal component (PC2), P_{annual} contributed the most, with an estimated coefficient of -0.884. For the third principal component (PC3), T_{rainy} contributed the most, and had an estimated coefficient of 0.660 (Table 2). Maps with the spatial distribution of each principal component were generated, but it was not possible to visualize a clear relationship between magnitude of the values and HLB presence (Figs. 2G, 2H, and 2I).

Table 2. Loading of the first three principal components (PC1, PC2, and PC3) from the principal component analysis using climate and orchard-related variables.

Variable[a]	PC1	PC2	PC3
A	0.244	-0.147	0.653
$N_{mandarin}$	-	-	-0.630

N_{orange}	-0.164	-	0.649
P_{annual}	-0.279	-0.884	0.238
P_{dry}	-0.739	-0.566	-
P_{rainy}	-	-0.872	0.297
T_{annual}	0.292	0.689	0.621
T_{dry}	0.383	0.716	0.537
T_{rainy}	0.237	0.666	0.660
W_{annual}	0.974	0.154	0.115
W_{dry}	0.883	0.219	0.364
W_{rainy}	0.974	0.109	-
Variance (%)	30.7	28.4	22.0
Cumulative var. (%)	30.7	59.1	81.1

[a] A = orchard area (m²); $N_{mandarin}$ = the log-transformed number of mandarin orchards; N_{orange} = log-transformed number of orange orchards; P_{annual} = mean daily precipitation (rainfall) in the year; P_{dry} = mean rainfall in the dry season (May to August); P_{rainy} = mean rainfall in the rainy season (September to April); T_{annual} = annual mean temperature; T_{dry} = mean temperature in the dry season; T_{rainy} = mean temperature in the rainy season; W_{annual} = annual windy speed; W_{dry} mean wind speed in the dry season; W_{rainy} = mean wind speed in the rainy season.

Using only the first three principal components and the spatial effect, the number of models evaluated was 16 (equivalent to 2^4). The model with the lowest WAIC (214.503) and a moderate LCPO value (6.35) was the model developed from PC1 and PC2 plus the spatial effect. The mean of the posterior distribution of the parameter for PC1 was positive, but the HPD interval contains zero indicating insufficient evidence of an effect of PC1 (Table 3). However, the mean of the posterior distribution of the parameter for PC2 was negative and its HPD interval does not contain zero, indicating sufficient evidence of an effect of PC2 (Table 3). The mean of the posterior distribution of the mixing parameter ϕ was slightly higher than that of the model with climate covariates but also indicating a partial effect of the spatial effect on HLB presence (Table 3). The across-grid mean and standard deviation of the posterior distribution of the spatial effect is depicted in Figs. 4A and 4B. Mean values ranged from -2.39 to 2.41, while standard deviation ranged from 0.35 to 1.65.

Table 3. Mean, median, standard deviation (SD), 95% highest posterior density (HPD) interval, and mode of model parameters and hyperparameters of the best model with principal components as covariates for citrus huanglongbing prevalence distribution in Minas Gerais, Brazil.

Parameters[b]	**Mean**	**Median**	**SD**	**HPD$_{lower}$**	**HPD$_{upper}$**	**Mode**
β_0	-0.396	-0.391	0.207	-0.808	0.006	-0.383
PC1	0.036	0.034	0.212	-0.379	0.455	0.029
[a] *PC2*	-0.895	-0.891	0.252	-1.394	-0.405	-0.882

Hyperparameters	Mean	Median	SD	HPD_{lower}	HPD_{upper}	Mode
τ	0.623	0.602	0.169	0.324	0.959	0.562
ϕ	0.172	0.128	0.148	0.002	0.482	0.033

[a] 95% HPD interval excludes 0.
[b] β_0 = model intercept; *PC1* = first principal component; *PC2* = second principal component; τ = variance; ϕ = the mixing parameter (spatial effect).

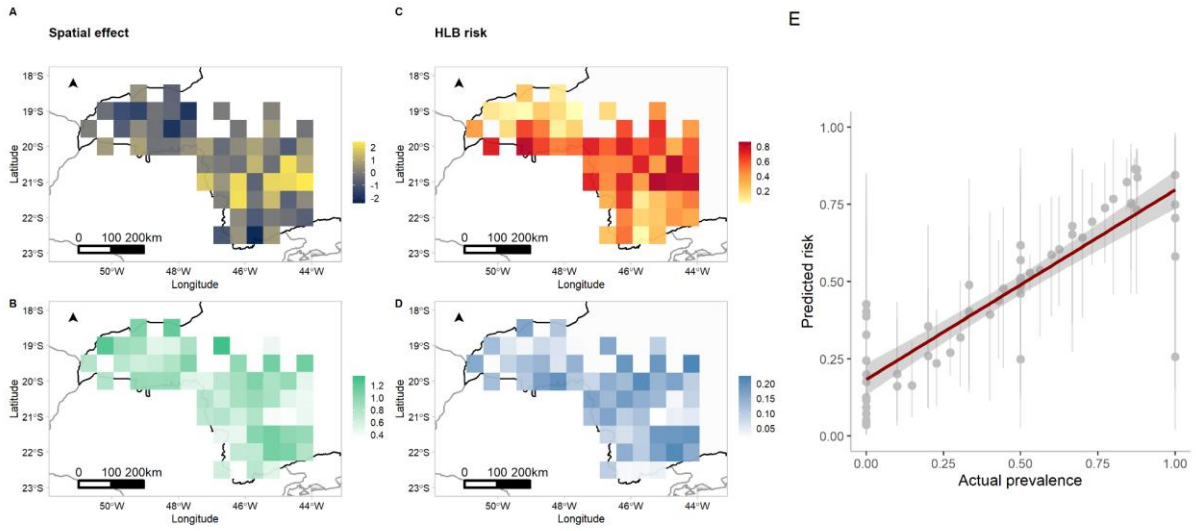

Fig. 4. Model with principal components and the spatial effect. Map (A) depicts the mean, and map (B) the standard deviation of the posterior distribution of the spatial effect. Map (C) depicts the posterior mean and map (D) the standard deviation of the spatial distribution of HLB prevalence (or risk) in Minas Gerais obtained using the model with the first and second principal components (PC1 and PC2) and the spatial effect as covariates. (E) shows the relationship between actual prevalence and the posterior mean (dots) and 95% highest posterior density (error bars) obtained with the principal components model.

The posterior mean of HLB prevalence in MG ranged from 0.035 to 0.865. The highest values are also in a region with a high density of orchards in the eastern region of the grid (Fig. 4C). Lower posterior probabilities were located in peripheral regions over western and southern Minas Gerais. The posterior standard deviation of the probabilities ranged from 0.019 to 0.223, and its spatial distribution was very similar to the model with climate covariates (Fig. 4D). Overall, there was a high correlation between the mean risk and actual HLB prevalence (Fig. 4E), in which the estimated coefficient of determination (r^2) was 0.75, which indicates high model precision. In fact, the mean HLB-risks predicted by both models presented here had a strong positive correlation (Pearson's r = 0.99).

DISCUSSION

Using two different modeling approaches (CC and PC), climate-driven models were successfully used to describe regional prevalence of HLB in MG, an important sweet-orange and mandarin-production region of Brazil. Among several climatic variables, we found evidence that the temperature during the dry season and annual wind speed were important in explaining the spatial distribution of HLB prevalence based on the CC model. In contrast, the PC model suggested temperature was a key factor and included mean annual and rainy season rainfall as important variables. Both models suggested a partial spatial effect, although the CC model accounted for more spatial effect than that reported for the PC model. Both models performed with high precision and provided comparable HLB predictions. Our findings are in general agreement with the current knowledge of factors that affect HLB epidemics (Antolinez et al. 2021; Carmo-Sousa et al. 2020; Martini et al. 2018; Gasparoto et al. 2012). Models suggested an inverse relationship between temperatures during the dry (winter) period and HLB development and spread, whereas wind speed was directly associated with a greater likelihood of HLB development. With the PC model, increased mean annual and rainy season rainfall favored an increased prevalence of HLB.

During May to August (autumn and winter months), minimal to no-host flushing occurs, and as a result *D. citri* densities are lower, which has been suggested to reduce the risk of HLB (Gutierrez and Ponti 2013; Lewis-Rosenblum et al. 2015; Teck et al. 2011; Hall and Albrigo 2007). The temperature for the long-term, 30-year climate normals for the winter across the area of study ranged between 15 and 22°C, more often 18°C or greater (data not shown). Multiple studies have identified mild (17-22°C) to hot (32-50°C) temperatures as detrimental to CLas replication and survival (Gasparoto et al. 2012; Razi et al. 2014; Lopes et al. 2009). In agreement with those reports, the results of our study suggest that mean winter temperatures are inversely related to prevalence of HLB in MG, suggesting that in regions where temperatures are higher during the winter, the conditions may be detrimental to CLas. However, recently, Lopes et al. (2017) reported temperatures below 15°C favored CLas titer in flushes on HLB-symptomatic trees, which had a lower frequency of occurrence across the region in our study. Also, the authors speculated drier periods were associated with host trees experiencing water deficiency, reducing tissue availability for *D. citri* feeding and reproduction, which corroborates the negative dry season effect on HLB prevalence reported in our model.

Long and short-distance dispersal of CLas is dependent on the dissemination of *D. citri* (Martini et al. 2018; Tomaseto et al. 2018; Antolinez et al. 2021; Lewis-Rosenblum et al. 2015).

Tomaseto et al. (2018) reported that at 27.14 ℃, approximately 50% of adult *D. citri* will initiate flight, although Martini et al. (2018) reported that a wider temperature range of 18 to 28 ℃ lengthened flight duration and distance of *D. citri*. Recently, a decline in flight and no long-distance flight initiation was observed at temperatures greater than 32 ℃ and 43 ℃, respectively (Antolinez et al. 2021). The findings are in agreement with our results in relation to the inverse effect of temperature on HLB prevalence. The decrease in temperature may act as a stimulus that influences psyllid flight responses, which is an important response in regards to long-distance dispersal of CLas.

Adult psyllids are responsible for the spreading CLas, whereas, *D. citri* nymphs have greater acquisition efficiency, and support more rapid CLas replication (George 2018; Ammar et al. 2016; Gottwald 2010). The voluntary and involuntary movement of adult psyllids is associated with wind speed and direction (Carmo-Sousa et al. 2020). The mean annual wind speed in MG ranged between 6.19 to 12.8 km/h. The increase of wind speed in wind tunnels has been suggested to be directly associated with the involuntary movement of psyllids (Martini et al. 2018). The previous results support our thesis that an increase in wind speed is directly associated with HLB prevalence, which is due to a greater probability of psyllids movement, and thus greater risk of HLB spread to, and establishment in new areas. Elucidating optimal wind speed for flight initiation and termination may assist in our understanding of *D. citri* movement, and the concomitant spread of CLas. Our study did not include a wind direction component, although Antolinez et al. (2021) suggested that the inclusion of wind direction is critical in understanding the effect of wind on flight orientation. Therefore, we can only speculate the contribution of wind direction to the partial spatial component of the CC model.

Although adult psyllids are capable of acquiring CLas, the acquisition time required is much longer, impacting temporal and spatial spread of the pathogen, and its resulting prevalence (Ammar et al. 2016). The reduced contribution of the partial spatial effect may be a result of adult *D. citri* movement predominating during the spring and summer months, i.e., during the rainy season in MG, resulting in greater pathogen dissemination through vector dispersal (Lewis-Rosenblum et al. 2015b; George 2018). However, a second explanation to the reduced spatial effect could be the long between-cell distance (100 km) used to create the neighborhood structure (spatial structure used as random effect) for modeling. If the grid-cells had smaller dimensions and a neighborhood structure was constructed using a lower distance threshold (<100 km), which would imply in a spatial dependence in smaller distances, maybe the spatial effect would have greater significance, since psyllids tend to disperse across distances less than 100 km.

Although the interpretation of effects due to individual variables during a modeling procedure when using PCs as covariates is challenging (Cendoya et al. 2020), the PC model allows sufficient insight that greater rainfall due to both the annual and rainy season period result in a greater risk of HLB. Based on the PCA model, warmer winters are associated with periods of reduced disease favorability, lowering the probability of HLB infection events, which is similar to the results of the CC model. But for the PC model, rainfall had more influence than temperature on HLB risk, which was due to the contributions of PC2.

Rainfall is associated with citrus host physiology, and in turn, with *D. citri* population dynamics. Greater populations of psyllid adults are associated with rainy-season summer periods where more flush tissue is available (Shimwela et al. 2018; Tsai et al. 2002; Lewis-Rosenblum et al. 2015; Gutierrez and Ponti 2013; Narouei-Khandan et al. 2016). In a multiple regression analysis, rainfall was associated with periods of host vegetative flushes, with a period of vegetative flushes in the rainy season and a second period of vegetative flushes in the dry season, and was significantly associated with favoring higher titer of CLas in symptomatic flushes (Lopes et al. 2017). Our results corroborate these findings, providing evidence of a greater risk for HLB associated with higher annual and rainy season rainfall. Rainfall should be included as a regional risk assessment parameter, and should improve the targeting and implementations of management strategies already in place.

Sampling and quantifying CLas titer in symptomatic trees (i.e., when flushes are present) or trapping psyllids in areas predicted to have a higher prevalence of HLB, would confirm the assumptions of our model. Furthermore, oviposition and nymph development occur exclusively on young flushes (Hall et al. 2008). The risk for CLas acquisition by nymphs is greater and more rapid compared to with adult *D. citri*, and as a result the probability for transmission to neighboring hosts increases as the *D. citri* nymphs mature (Pelz-Stelinski et al. 2010; George 2018; Ammar et al. 2016; Gottwald 2010).

We speculate that areas closer to HLB orchards will be more likely to become infected compared to those farther away, suggesting a clustering or coalescence of HLB events, although the spatial effect was only partial. In this study, citrus orchards adjacent to SP would have a high probability of having trees with HLB, in contrast to areas in the northwest and south of MG, where there were lower posterior mean probabilities. As CLas was first reported in SP, this result could be expected (Coletta-Filho et al. 2004; do Carmo Teixeira et al. 2005). Furthermore, Alves et al. (2020) indicated a greater spread of HLB in the areas bordering SP compared to those areas in central and south MG, although, larger populations of CLas positive

psyllids were present in southwestern and northern SP, compared to the state border between MG and SP (Wulff et al. 2020).

The two models (CC and PC) we have presented could motivate targeted monitoring in regions with a higher HLB risk. Furthermore, scouting in higher-risk regions for *Citrus* spp., in backyards, public gardens, pasture, and forests may also contribute to mitigating HLB through eradicating and replacing CLas-infected trees (Bassanezi et al. 2013). Management strategies should consider climatic effects when timing vector control intervention in HLB favorable seasons and regions, in order to minimize dispersal of CLas between and within orchards, thereby enhancing area-wide management.

ACKNOWLEDGMENTS

K.S.A. acknowledges CAPES for providing a graduate scholarship. EMD is thankful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for providing a research fellowship. Climatic data were obtained from the NASA Langley Research Center POWER Project funded through the NASA Earth Science Directorate Applied Science Program.

REFERENCES

Alves, K. S., do Carmo, L. H. M., and Del Ponte, E. M. 2020. Spatiotemporal spread of huanglongbing in commercial citrus orchards of Minas Gerais, Brazil. Trop. Plant Pathol. 45:668–679.

Ammar, E.-D., Ramos, J. E., Hall, D. G., Dawson, W. O., and Shatters, R. G. 2016. Acquisition, Replication and Inoculation of *Candidatus* Liberibacter asiaticus following Various Acquisition Periods on Huanglongbing-Infected Citrus by Nymphs and Adults of the Asian Citrus Psyllid ed. Murad Ghanim. PLOS ONE. 11:e0159594 Available at: https://dx.plos.org/10.1371/journal.pone.0159594 [Accessed October 2, 2020].

Antolinez, C. A., Moyneur, T., Martini, X., and Rivera, M. J. 2021. High Temperatures Decrease the Flight Capacity of *Diaphorina citri* Kuwayama (Hemiptera: Liviidae). Insects. 12:394 Available at: https://www.mdpi.com/2075-4450/12/5/394 [Accessed May 4, 2021].

Bassanezi, R. B., Busato, L. A., Filho, A. B., Amorim, L., and Gottwald, T. R. 2005. Preliminary spatial pattern analysis of Huanglongbing in São Paulo, Brazil. Int. Organ. Citrus Virol. Conf. Proc. 1957-2010. 16 Available at: https://escholarship.org/uc/item/76s629f9#author [Accessed April 12, 2021].

Bassanezi, R. B., Lopes, S. A., de Miranda, M. P., Wulff, N. A., Volpe, H. X. L., and Ayres, A. J. 2020. Overview of citrus huanglongbing spread and management strategies in Brazil. Trop. Plant Pathol. 45:251–264.

Bassanezi, R. B., Montesino, L. H., Gimenes-Fernandes, N., Yamamoto, P. T., Gottwald, T. R., Amorim, L., et al. 2013. Efficacy of Area-Wide Inoculum Reduction and Vector Control on Temporal Progress of Huanglongbing in Young Sweet Orange Plantings. Plant Dis. 97:789–796.

Besag, J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. J. R. Stat. Soc. Ser. B Methodol. 36:192–225.

Besag, J., York, J., and Mollié, A. 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 43:1–20.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. 2013. *Applied spatial data analysis with R, Second edition*. Springer, NY. Available at: https://asdar-book.org/.

Bové, J. M. 2006. Huanglongbing: a destructive, newly-emerging, century-old disease of citrus. J. Plant Pathol. 88:7–37 Available at: https://www.jstor.org/stable/41998278 [Accessed October 2, 2020].

do Carmo Teixeira, D., Saillard, C., Eveillard, S., Danet, J. L., Costa, P. I. da, Ayres, A. J., et al. 2005. '*Candidatus* Liberibacter americanus', associated with citrus huanglongbing (greening disease) in São Paulo State, Brazil. Int. J. Syst. Evol. Microbiol. 55:1857–1862 Available at: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.63677-0 [Accessed October 2, 2020].

Carmo-Sousa, M., Cortés, M. T. B., and Lopes, J. R. S. 2020. Understanding psyllid transmission of *Candidatus* Liberibacter as a basis for managing huanglongbing. Trop. Plant Pathol. Available at: https://doi.org/10.1007/s40858-020-00386-1 [Accessed October 2, 2020].

Carvalho, S. A. D., Girardi, E. A., Mourão Filho, F. D. A. A., Ferrarezi, R. S., & Coletta

Filho, H. D. 2019. Advances in citrus propagation in Brazil. Revista Brasileira de Fruticultura 41 (6). http://dx.doi.org/10.1590/ 0100-29452019422

Cendoya, M., Martínez-Minaya, J., Dalmau, V., Ferrer, A., Saponari, M., Conesa, D., et al. 2020. Spatial Bayesian Modeling Applied to the Surveys of *Xylella fastidiosa* in Alicante (Spain) and Apulia (Italy). Front. Plant Sci. 11 Available at: https://www.frontiersin.org/articles/10.3389/fpls.2020.01204/full [Accessed August 24, 2020].

Coletta-Filho, H. D., Targon, M. L. P. N., Takita, M. A., De Negri, J. D., Pompeu, J., Machado, M. A., et al. 2004. First Report of the Causal Agent of Huanglongbing ("*Candidatus* Liberibacter asiaticus") in Brazil. Plant Dis. 88:1382–1382 Available at: https://apsjournals.apsnet.org/doi/abs/10.1094/PDIS.2004.88.12.1382C [Accessed October 2, 2020].

Craig, A. P., Cunniffe, N. J., Parry, M., Laranjeira, F. F., and Gilligan, C. A. 2018. Grower and regulator conflict in management of the citrus disease Huanglongbing in Brazil: A modelling study. J. Appl. Ecol. 55:1956–1965 Available at: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.13122 [Accessed October 18, 2020].

Fundecitrus. 2019. Fundecitrus. Available at: https://www.fundecitrus.com.br/pdf/levantamentos/levantamento-doencas-2019.pdf [Accessed October 2, 2020].

Gasparoto, M. C. G., Coletta-Filho, H. D., Bassanezi, R. B., Lopes, S. A., Lourenço, S. A., and Amorim, L. 2012. Influence of temperature on infection and establishment of '*Candidatus* Liberibacter americanus' and '*Candidatus* Liberibacter asiaticus' in citrus plants. Plant Pathol. 61:658–664 Available at: https://bsppjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3059.2011.02569.x [Accessed October 25, 2020].

Gasparoto, M. C. G., Hau, B., Bassanezi, R. B., Rodrigues, J. C., and Amorim, L. 2018. Spatiotemporal dynamics of citrus huanglongbing spread: a case study. Plant Pathol. 67:1621–1628.

George, J. 2018. Prolonged phloem ingestion by *Diaphorina citri* nymphs compared to adults is correlated with increased acquisition of citrus greening pathogen. Sci. Rep. :11.

Gottwald, T., Irey, M., Gast, T., Parnell, S., Taylor, E., Hilf, M., et al. 2010. Spatio-temporal Analysis of an HLB Epidemic in Florida and Implications for Spread.

Gottwald, T. R. 2010. Current epidemiological understanding of citrus huanglongbing. Annu. Rev. Phytopathol. :119–139 Available at: https://pubag.nal.usda.gov/download/47697/PDF [Accessed October 10, 2020].

Gutierrez, A. P., and Ponti, L. 2013. Prospective Analysis of the Geographic Distribution and Relative Abundance of Asian Citrus Psyllid (Hemiptera: Liviidae) and Citrus Greening Disease in North America and the Mediterranean Basin. Fla. Entomol. 96:1375–1391.

Hall, D. G., and Albrigo, L. G. 2007. Estimating the Relative Abundance of Flush Shoots in Citrus with Implications on Monitoring Insects Associated with Flush. HortScience. 42:364–368 Available at: https://journals.ashs.org/view/journals/hortsci/42/2/article-p364.xml [Accessed May 4, 2021].

Hall, D. G., Hentz, M. G., and Adair, R. C., Jr. 2008. Population Ecology and Phenology of *Diaphorina citri* (Hemiptera: Psyllidae) in Two Florida Citrus Groves. Environ. Entomol. 37:914–924 Available at: https://doi.org/10.1093/ee/37.4.914 [Accessed May 9, 2021].

Instituto Brasileiro De Geografia E Estatística – IBGE. 2018. Levantamento Sistemático da Produção Agrícola. Disponível em: https://www.ibge.gov.br/estatisticas-novoportal/economicas/agricultura-e-pecuaria/9201-levantamento-sistematico-da-producao-agricola.html?et=resultados.

Jagoueix, S., Bové, J.-M., and Garnier, M. 1994. The Phloem-Limited Bacterium of Greening Disease of Citrus Is a Member of the a Subdivision of the Proteobacteria. Int. J. Syst. Bacteriol. :379–386.

Leal, R. M., Barbosa, J. C., Costa, M. G., Belasque Junior, J., Yamamoto, P. T., and Dragone, J. 2010. Distribuição espacial de Huanglongbing (Greening) em citros utilizando a geoestatística. Rev. Bras. Frutic. 32:808–818.

Lewis-Rosenblum, H., Martini, X., Tiwari, S., and Stelinski, L. L. 2015a. Seasonal Movement Patterns and Long-Range Dispersal of Asian Citrus Psyllid in Florida Citrus. J. Econ. Entomol. 108:3–10 Available at: https://doi.org/10.1093/jee/tou008 [Accessed April 12, 2021].

Lopes, S. A., Bassanezi, R. B., Jr, J. B., and Yamamoto, P. T. 2008. Management of Citrus

Huanglongbing in the State of São Paulo-Brazil. Available at: https://www.fftc.org.tw/htmlarea_file/library/2011071 2174730/eb609.pdf.

Lopes, S. A., Bertolini, E., Frare, G. F., Martins, E. C., Wulff, N. A., Teixeira, D. C., et al. 2009. Graft Transmission Efficiencies and Multiplication of '*Candidatus* Liberibacter americanus' and ' *Ca.* Liberibacter asiaticus' in Citrus Plants. Phytopathology®. 99:301–306 Available at: https:// apsjournals.apsnet.org/doi/10.1094/PHYTO-99-3-0301 [Accessed October 25, 2020].

Lopes, S. A., Luiz, F. Q. B. F., Oliveira, H. T., Cifuentes-Arenas, J. C., and Raiol-Junior, L. L. 2017. Seasonal Variation of ' *Candidatus* Liberibacter asiaticus' Titers in New Shoots of Citrus in Distinct Environments. Plant Dis. 101:583–590 Available at: https://apsjournals.apsnet.org/doi/10.1094/PDIS-06-16-0859-RE [Accessed May 8, 2021].

Madden, L. V., Hughes, G., and van den Bosch, F. 2007. *The Study of Plant Disease Epidemics*. St. Paul: APS Press. Available at: https://apsjournals.apsnet.org/doi/book/10.1094/9780890545058 [Accessed March 8, 2019].

Martini, X., Rivera, M., Hoyte, A., Sétamou, M., and Stelinski, L. 2018. Effects of Wind, Temperature, and Barometric Pressure on Asian Citrus Psyllid (Hemiptera: Liviidae) flight behavior. J. Econ. Entomol. 111:8.

Narouei-Khandan, H. A., Halbert, S. E., Worner, S. P., and van Bruggen, A. H. C. 2016. Global climate suitability of citrus huanglongbing and its vector, the Asian citrus psyllid, using two correlative species distribution modeling approaches, with emphasis on the USA. Eur. J. Plant Pathol. 144:655–670 Available at: http://link.springer.com/10.1007/s10658-015-0804-7 [Accessed May 8, 2021].

Pelz-Stelinski, K. S., Brlansky, R. H., Ebert, T. A., and Rogers, M. E. 2010. Transmission Parameters for *Candidatus* Liberibacter asiaticus by Asian Citrus Psyllid (Hemiptera: Psyllidae). J. Econ. Entomol. 103:1531–1541 Available at: https://academic.oup.com/jee/article-lookup/doi/10.1603/EC10123 [Accessed May 9, 2021].

Pettit, L. I. 1990. The Conditional Predictive Ordinate for the Normal Distribution. J. R. Stat. Soc. Ser. B Methodol. 52:175–184.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Razi, M. F., Keremane, M. L., Ramadugu, C., Roose, M., Khan, I. A., and Lee, R. F. 2014. Detection of Citrus Huanglongbing-Associated '*Candidatus* Liberibacter asiaticus' in Citrus and *Diaphorina citri* in Pakistan, Seasonal Variability, and Implications for Disease Management. Phytopathology®. 104:257–268 Available at: https://apsjournals.apsnet.org/doi/10.1094/PHYTO-08-13-0224-R [Accessed May 4, 2021].

Revelle W, 2020. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University.

Rue, H., Martino, S., and Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71:319–392.

de Sá Júnior, A., de Carvalho, L. G., da Silva, F. F., and de Carvalho Alves, M. 2012. Application of the Köppen classification for climatic zoning in the state of Minas Gerais, Brazil. Theor. Appl. Climatol. 108:1–7 Available at: https://doi.org/10.1007/s00704-011-0507-8 [Accessed November 16, 2020].

Passos, O. S., da Silva Souza, J., Costa Bastos, D., Girardi, E. A., de Lima Gurgel, F., Bastos Garcia, M. V., de Oliveira, R. P. and dos Santos Soares Filho, W. Citrus Industry in Brazil with Emphasis on Tropical Areas in Citrus-Health Benefits and Production Technology. IntechOpen, 2018. Available at: DOI: 10.5772/intechopen.80213.

Shen, W., Halbert, S. E., Dickstein, E., Manjunath, K. L., Shimwela, M. M., and Bruggen, A. H. C. van. 2013. Occurrence and in-grove distribution of citrus huanglongbing in north central Florida. J. Plant Pathol. 95:361–371.

Shimwela, M. M., Schubert, T. S., Albritton, M., Halbert, S. E., Jones, D. J., Sun, X., et al. 2018. Regional Spatial-Temporal Spread of Citrus Huanglongbing Is Affected by Rain in Florida. Phytopathology®. 108:1420–1428 Available at: https://apsjournals.apsnet.org/doi/10.1094/PHYTO-03-18-0088-R [Accessed October 2, 2020].

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. 2017. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. Stat.

Sci. 32:1–28.

Sparks, A. H. 2018. nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. J. Open Source Softw. 3:1035.

Teck, S. L. C., Fatimah, A., Beattie, A., Heng, R. K. J., and King, W. S. 2011. Seasonal Population Dynamics of the Asian Citrus Psyllid, *Diaphorina citri* Kuwayama in Sarawak. Am. J. Agric. Biol. Sci. 6:527–535 Available at: https://thescipub.com/abstract/ajabssp.2011.527.535 [Accessed May 4, 2021].

Tomaseto, A. F., Miranda, M. P., Moral, R. A., Lara, I. A. R. de, Fereres, A., and Lopes, J. R. S. 2018. Environmental conditions for *Diaphorina citri* Kuwayama (Hemiptera: Liviidae) take-off. J. Appl. Entomol. 142:104–113 Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/jen.12418 [Accessed April 12, 2021].

Tsai, J. H., Wang, J.-J., and Liu, Y.-H. 2002. Seasonal Abundance of the Asian Citrus Psyllid, *Diaphorina citri* (Homoptera: Psyllidae) in Southern Florida. Fla. Entomol. :446–446 Available at: https://journals.flvc.org/flaent/article/view/75122 [Accessed May 8, 2021].

Van Vuuren, S. P. and Da Graça, J. V. 1993. Variable transmission of African greening to sweet orange. Pages 264-268 in Proc. 12th Conference of the International Organization of Citrus Virologists. University of California, Riverside. Available at: https://swfrec.ifas.ufl.edu/hlb/ database/pdf/00000503.pdf

Watanabe, S. 2010. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. J. Mach. Learn. Res. 11:3571–3594.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. 2019. Welcome to the tidyverse. J. Open Source Softw. 4:1686.

Wulff, N. A., Daniel, B., Sassi, R. S., Moreira, A. S., Bassanezi, R. B., Sala, I., et al. 2020. Incidence of *Diaphorina citri* Carrying *Candidatus* Liberibacter asiaticus in Brazil's Citrus Belt. Insects. 11:672 Available at: https://www.mdpi.com/2075-4450/11/10/672 [Accessed May 8, 2021].

Xie, Y., Allaire, J. J., and Grolemund, G. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC. Available at: https://bookdown.org/yihui/rmarkdown.

CHAPTER 2: Modeling the effect of El Niño Southern Oscillation on the onset of soybean rust in Southern Brazil

This manuscript has been submitted for publication.
Formatting follows the journal requirements.

ABSTRACT

The onset of soybean rust (SBR) epidemics, caused by the fungus *Phakopsora pachyrhizi*, in Brazil is mainly driven by the availability of inoculum and weather conditions that prevail during the pre- or early-season. The effect of the El Niño-southern oscillation (ENSO) on weather patterns of certain regions is well known. In Southern Brazil, rainfall levels are increased above normal in spring and summer during the warm (El Niño) phase, which in turn may affect epidemic patterns in summer crops. In this study, data on the first report of SBR in a municipality recorded during 17 growing seasons (2004/05 to 2020/21) were gathered from the Brazilian anti-rust consortium (*Consórcio Antiferrugem*). The Oceanic Niño Index (ONI), or the average anomaly on the sea surface temperature in the El Niño 3.4 region for three consecutive months, was gathered from two trimesters prior to the growing season [January, February, and March (JFM) and May, June, and July (MJJ)] and one trimester from within-growing season [October, November, and December (OND)]. A spatially structured Bayesian Cox proportional model suggested that the increase in the probability of early SBR onset is associated with the increase in ONI from MJJ and OND. No association was found for the ONI from JFM on disease onset. Our results provide novel insights into the large-scale risk of SBR outbreaks in Southern Brazil, whose knowledge may be useful for early-planning of disease management strategies.

Keywords: *Phakopsora pachyrhizi*, outbreaks, teleconnections, survival analysis, climate

INTRODUCTION

Soybean rust (SBR), a fungal disease caused by *Phakopsora pachyrhiz*i, was first reported in South America in 2001 (Yorinori et al. 2005). The disease has spread to the major soybean (*Glycine max*) areas in the Americas and is currently a major concern for soybean growers due to its potential to severely reduce crop yield, particularly in Brazil where conditions are

favorable for epidemic onset and development (Del Ponte et al. 2006c; Dalla Lana et al. 2015; Godoy et al. 2016). Since the early years of dealing with the disease, major actions were taken to understand the disease, prevent its further spread, predict disease risk, and establish effective control measures, more notably in Brazil (Godoy et al. 2016) and the United States (Allen et al. 2014; Sikora et al. 2014). In Brazil, a national "anti-rust" consortium named "Consórcio Antiferrugem"' was established in 2004, when efforts focused on monitoring the seasonal disease spread across the country as well as testing fungicides for disease control (Godoy et al. 2016; Dalla Lana et al. 2018; Barro et al. 2021; Machado et al. 2021). The initiative has been responsible for collecting, curating, and sharing data on the first SBR detections at the municipality level every growing season (http://www.consorcioantiferrugem.net).

In general, the main soybean growing season in Brazil starts with the onset of the Spring season. In the tropics, sowing time is strongly dependent on the onset of the rainy season (usually by mid-September). Hence, delays in sowing time are not uncommon due to early-season droughts, which are responsible for moving sowing one or two months further. Prior to the introduction of the soybean rust pathogen into Brazil, winter (June-September) soybean cropping was a common practice targeting seed production. However, these crops were found to serve as an efficient 'green bridge' that allowed the movement of soybean rust inoculum across seasons (Godoy et al. 2016). Since 2005, it has become mandatory to establish a 90-day soybean-free period in 14 Brazilian states, with the goal of reducing airborne inoculum during the early season (Godoy et al. 2016). Indeed, the sowing period is allowed from mid-September up to December 31st in six states. The adoption of a soybean-free period has helped to decrease the disease risk (Godoy et al. 2016), but volunteer soybean plants and alternative hosts, which are impossible to eliminate completely, allow the fungus to survive between seasons (Yorinori et al. 2005). Once established, SBR epidemics may spread rapidly throughout the country/continent via transportation of the rust spores by air currents (Isard et al. 2011; Pan et al. 2006), which favors disease occurrence in every season (Del Ponte et al. 2006a).

Studies to elucidate factors (mainly weather-based) affecting the components of the disease cycle, as well as of observed epidemics, have been conducted to deepen the understanding of the fungal biology and epidemiology, as well as to develop and improve disease warning systems (Del Ponte et al. 2011; Del Ponte and Esker 2008; Del Ponte et al. 2006a; Beruski et al. 2020; Marchetti 1976; Koch and Hoppe 1987; Alves et al. 2011). The most important weather-related variables that drive the development of SBR epidemics include rainfall (frequency and amount), leaf wetness duration, and temperature, which are critical for

infection and further inoculum production during the successive cycles (Del Ponte et al. 2006c, 2006b; Del Ponte and Esker 2008).

In Brazil, the seasonal weather of some regions is under the influence of the El Niño–Southern Oscillation (ENSO) (Cirino et al. 2015). ENSO phases (El Niño, neutral, or La Niña) are determined by the Oceanic Niño Index (ONI), which is calculated based on anomalies in the normal sea surface temperature (SST) in the El Niño 3.4 region in the central Pacific, a region known to influence climate at the global scale (Grimm 2003). In South America, the warm phase (El Niño) usually leads to an increase in rainfall levels in the south of the country (subtropics) and droughts in the tropics during the spring and summer months. The pattern is reversed during the cold phase (La Niña) (Grimm 2003; Cirino et al. 2015; Cai et al. 2020).

The effects of the ENSO phases, or the ONI directly, have been investigated previously for human and plant diseases. Studies have linked El Niño events to outbreaks of malaria, cholera, plague, and dengue (Anyamba et al. 2019; Rosenzweig et al. 2001; Kovats et al. 2003). Fewer examples exist linking plant diseases and ENSO. In the United States, spectral and cross-spectral analysis was used to link a few oceanic indices (ONI, Pacific-North American pattern, or the North Atlantic Oscillation) to fluctuations of Fusarium head blight of wheat (Kriss et al. 2012) and wheat rust (Scherm and Yang 1995). In Brazil, the apparent infection rates of coffee leaf rust epidemics were predicted to be greatest during El Niño seasons in most locations of the subtropical region (Hinnah et al. 2020). For soybean rust, three studies conducted in Brazil used weather-based models, which were linked to a long series of historical weather and crop models, to predict SBR risk and investigate its association with the ENSO phases (Del Ponte et al. 2011; Radons et al. 2021; Fattori et al. 2021). The effect of SST in the El Niño 3.4 region on the SBR final seasonal prevalence (cumulative number of reports) was investigated for two states: Paraná and Mato Grosso during 11 growing seasons (2004 to 2014) (Minchio et al. 2018).

Because ENSO is known to trigger early rainfall in Southern Brazil (Grimm 2003; Cirino et al. 2015; Cai et al. 2020) and soybean rust epidemics are highly dependent on rainfall (Del Ponte et al. 2006c), we hypothesize that ENSO also plays an important role in the time to outbreak onset on a regional basis. Therefore, in this study, we gathered data from soybean monitoring in Southern Brazil from 2005 to 2021 growing seasons to model the effect of anomalies in the SST (ONI) on the timing of regional epidemic onset.

MATERIAL AND METHODS

Data on soybean rust presence

The data on SBR onset in commercial fields in Southern Brazil were provided by the Brazilian national anti-rust consortium. The monitoring of SBR in Brazilian national territory is1 done by agricultural specialists from state plant health defense agencies. Whenever soybean rust-like symptoms are spotted in commercial fields, samples are sent to an official laboratory, which is responsible for the confirmation and reporting to the national anti-rust consortium. The disease occurrence is given at the municipality level, therefore, whenever SBR is found in a commercial soybean field, the respective municipality is set as SBR-positive. The frequency in which fields are monitored is not provided. Data on the date of SBR occurrence within 17 growing seasons (2004/05 to 2020/21) were gathered for the three southernmost states (Paraná, Santa Catarina, and Rio Grande do Sul) of Brazil. Data on the planting date of the commercial field in which SBR was found were also retrieved. The time for disease onset was calculated as the difference between occurrence and planting date. Hereafter the notation for growing seasons displays the year in which the season has ended or harvested, e.g. the 2004/05 growing season will be denoted as 2005.

Data on Oceanic Niño Index (ONI)

Data ranging from 2004 to 2021 on the Oceanic Niño Index (ONI) were obtained from the National Oceanic Atmospheric Administration (NOAA) website (NOAA (National Oceanic and Atmospheric Administration) 2022). ONI is a measurement of the ENSO, which tracks the anomaly on the sea surface temperature (SST) at an area in the east-central tropical Pacific Ocean located specifically at the 5ºN-5ºS latitude and 120º-170ºW longitude (Huang et al. 2017). ONI is calculated as a 3-month running average of the anomaly on SST from its climatic normal (30 years average, updated every 5 years). ONI data were retrieved from three trimesters: January, February, and March (ONI_{JFM}, prior to the growing season); May, June, and July (ONI_{MJJ}, prior to the growing season); and October, November, and December (ONI_{OND}, within the growing season).

The effect of ONI on the disease onset

The effect of ONI_{JFM}, ONI_{MJJ}, and ONI_{OND} on the disease onset was estimated through survival analysis using a Bayesian Cox proportional hazards (CPM) model. Survival analysis estimates

the probability of a certain event occurring (here is the SBR occurrence after the planting date). In this sense, survival analysis is used to estimate survival functions (equation 1).

$$S(t) = P(T > t) = 1 - F(t) \quad (4)$$

Where T is the random variable that measures the time to event, or in our case, the time to SBR report; the survival function $S(t)$ is the probability that T is higher than a given time t, therefore $S(t) = P(T > t)$. It gives the probability of disease occurrence for each time t, or yet, the probability of no occurrence of an outbreak over time. $F(t)$ is the cumulative probability distribution function of T. The distribution of $S(t)$ along all values of t gives the survival curve.

The hazard function $h(t)$ gives the conditional failure rate, which is defined as the probability of failure (or disease occurrence) in a small-time interval (Lee and Wang 2003). Its mathematical definition is given by equation 2, where $f(t)$ is the probability density function of T:

$$h(t) = \frac{f(t)}{S(t)} \quad (2)$$

The Cox proportional hazard (CPM) model is a product of the baseline hazard function $h_0(t)$, and a term accounting for a set of covariates x, which effects are regulated by their associated coefficients β.

$$h(t) = h_0(t)exp(\beta x) \quad (6)$$

The CPM was implemented under the Bayesian framework using the Integrated Nested Laplace Approximation (INLA). Instead of Markov chain Monte Carlo methods for computing the joint posterior distribution of model parameters, which is computationally expensive, INLA focuses on obtaining individual posterior marginals and on models that can be expressed as Gaussian Markov random fields, reducing computational expensiveness (Rue et al. 2009; Rue and Held 2005; Gómez-Rubio 2020). Bayesian inference using INLA can be performed in R using the INLA package (Lindgren and Rue 2015).

The CPM is implemented in INLA as shown in Equation 7, where η is a linear predictor on some covariates (fixed and random effects).

$$h(t) = h_0(t)exp(\eta) \quad (7)$$

The baseline hazard function is modeled as $h0(t) = exp(b_l); t \in (s_{l-1}, s_l], l = 1, \ldots, L$, therefore the baseline hazard is a piecewise constant function, where a random walk

prior is given to b_l, in which $b_l - b_{l-1} \sim N(0, \delta)$, where δ is the precision for the baseline hazard.

The fitted models (one model for each trimester) included ONI as fixed effects, the growing seasons as random effects, and the vector of municipalities was included a spatially structured random effect, in which the linear predictor η model would become a reparametrized Besag-York-Mollié model 194 (BYM-model) (Besag et al. 1991), taking into account the spatial dependence of neighboring municipalities. This type of modeling framework has been used to model the spatial distribution of Huanglongbing of citrus (Alves et al. 2021) and *X. fastidiosa* in Spain and Italy (Cendoya et al. 2020). The general form of the model is given by:

$$\eta_{ij} = \beta_0 + X_i\beta_m + Z_j + \frac{1}{\tau}\left(\sqrt{1-\phi}u_i + \sqrt{\phi}v_i^*\right) \quad (8)$$

Where β_0 is the overall intercept, β_m is the vector coefficients for the m covariates or predictors, X_i is the vector of covariates, Z_j is the vector of random effects of each j season, u_i is a zero-mean Gaussian with a precision matrix representing the unstructured random effect, v_i^* is the random effects scaled spatial component that accounts for the similarities of the neighboring municipalities (Simpson et al. 2017), and ϕ ($0 \leq \phi \leq 1$) is a mixing parameter which gives the fraction of the variance explained by the spatial structure, being directly proportional to the effect of the spatial structure. We assumed a distance of 200 km (maximum distance to consider two municipalities as neighbors) for building the spatial structure of random effects used in the models. The priors for all models used in this study were:

$$P(\beta_0) \propto 1$$
$$\beta_m \sim N(\mu = 0, \tau = 10^{-3}), m = 1, \ldots, N_\beta$$
$$Z_j \sim N(0.001, \lambda)$$
$$b_l - b_{l-1} \sim N(0, \delta)$$
$$\delta = 0.0001$$
$$\tau \sim PCprior(0.5/0.31, 0.01),$$
$$\phi \sim PCprior(0.5, 2/3)$$
$$\lambda \sim Gamma(0.001, 0.001)$$

Where PCprior(•) is the penalized complexity priors for the hyperparameters of the spatial component (Besag et al. 1991; Besag 1974; Simpson et al. 2017; Cendoya et al. 2020; Alves et al. 2021).

RESULTS

Summary of data on soybean rust onset

A total of 2,441 occurrences were recorded at the municipality level from 2005 to 2021 growing seasons. The state of Paraná had the highest number of occurrences (n = 1,239), followed by Rio Grande do Sul (n = 969) and Santa Catarina (n = 233). These differences are expected since these states have different numbers of soybean-growing municipalities, from which 278, 269, and 87 municipalities, respectively, have recorded at least one SBR occurrence. The growing season with fewer reported occurrences was 2012 (Fig. 1), with only 27 municipalities with SBR occurrences, while 2006 had the highest, with a total of 215 entries. The earliest first SBR occurrence in the growing season was in the 2010 growing season (Fig. 1) when SBR was detected as early as September (of 2009) when soybean plants were still at the vegetative stage (data not shown). On the other hand, the latest first SBR occurrence was recorded in the 2014 growing season (Fig. 1), when SBR symptoms were found in late December (of 2013) when soybean plants were at early reproductive stages (data not shown).

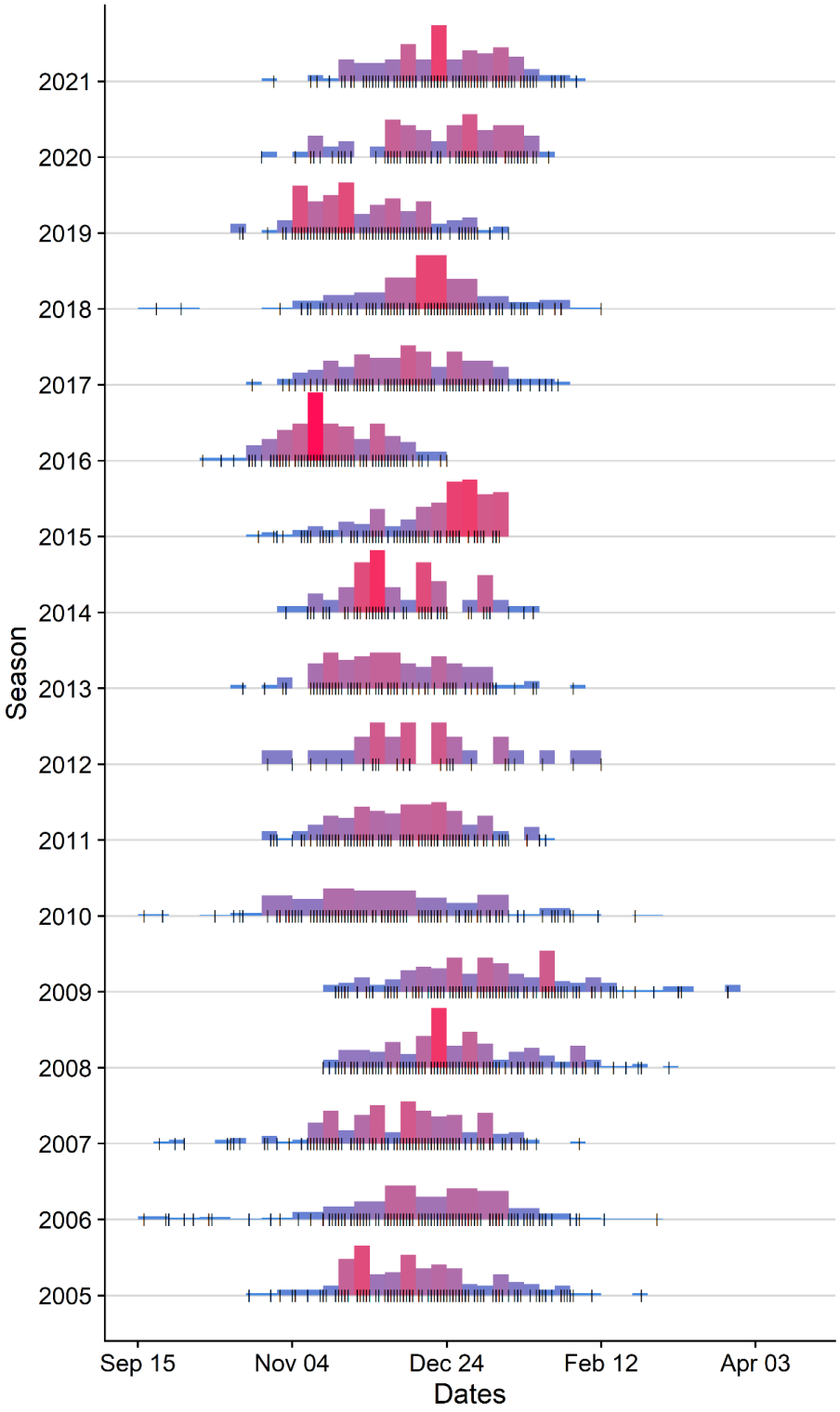

Fig. 1: Distribution of date of soybean rust (SBR) occurrences in Southern Brazil from 2005 to 2021 growing seasons. Vertical lines represent a unique date in which SBR was found in a commercial field.

Data on ONI

The ONI values of each trimester (JFM, MJJ, and OND) for each soybean growing season are depicted in Fig. 2. The highest value of ONI registered among these trimesters was 2.6° C in the OND trimester of the 2016 growing season (which refers to OND accessed in the year of 2015). The lower ONI value (-1.6°C) was also obtained from OND but for the 2011 growing season. The distribution of ONI values for each semester is depicted in Fig. 3. In general, the highest variation in ONI was obtained from the OND trimester, with a standard deviation (SD) of 1.15°C, followed by JFM (SD= 0.94°C) and MJJ (SD=0.47°C). The mean ONI values for each trimester were -0.05, -0.01, and -0.04, for JFM, MJJ, and OND, respectively.

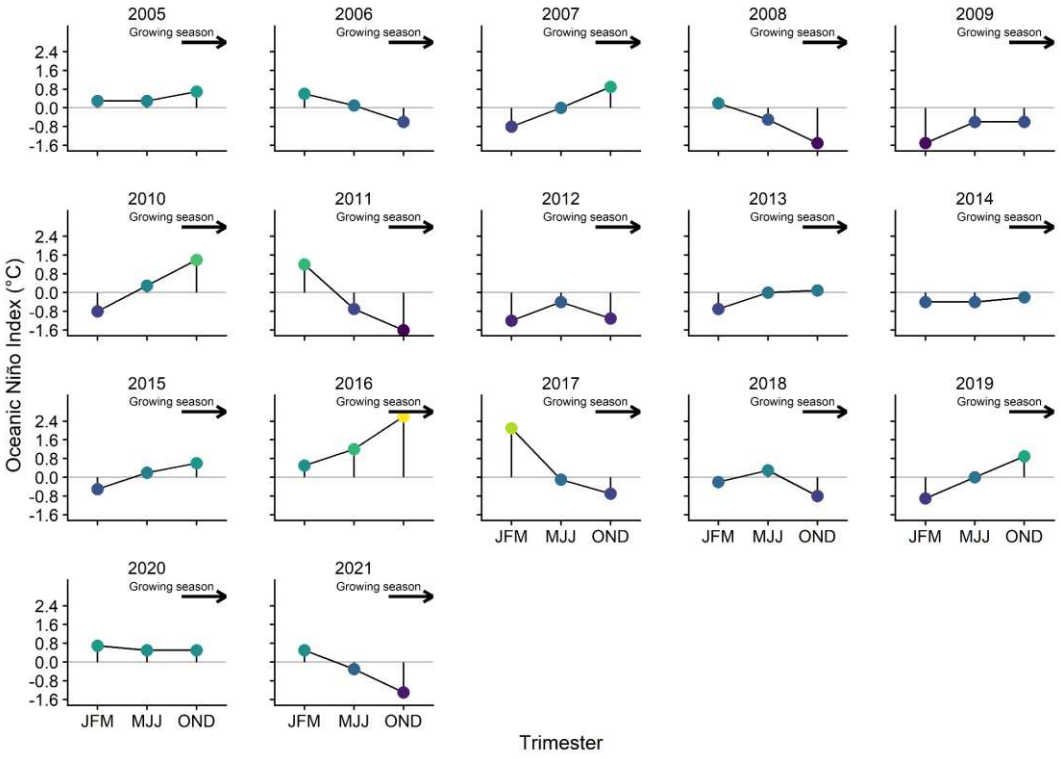

Fig. 2. Oceanic Niño Index (ONI) of the January, February, and March (JFM), May, June, and July (MJJ), and October, November, and December (OND) trimesters from 2005 to 2021 growing seasons. JFM and MJJ refer to trimesters before the beginning of the soybean growing season in Brazil.

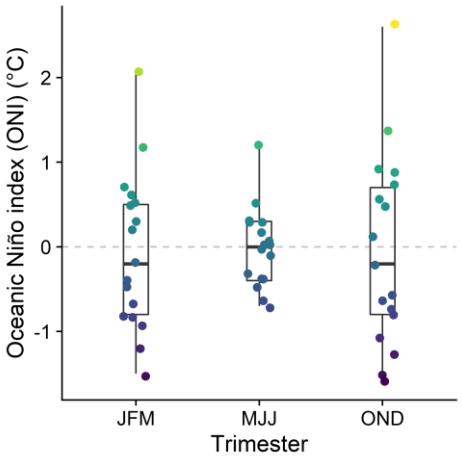

Fig. 3. Boxplot of the distribution of Oceanic Niño Index (ONI) of the January, February, and March (JFM), May, June, and July (MJJ), and October, November, and December (OND) trimesters from 2005 to 2021 growing seasons. JFM and MJJ refer to trimesters before the beginning of the soybean growing season in Brazil.

The effect of ONI on the disease onset time

There was not enough evidence to reject the null hypothesis that ONI_{JFM} is associated with the time of SBR onset in Southern Brazilian commercial fields of soybean in the following growing season (95% highest posterior density interval of its associated coefficient contained zero; Table 4). There was, however, evidence of the association of the ONI_{MJJ} and ONI_{OND} on the timing of SBR outbreaks (Table 4). Both models presented positive posterior means, which indicates positive associations of the ONI from those two periods, i.e. the increase in ONI in these trimesters is associated with the increase in the probability of early onsets of SBR epidemics in Southern Brazil. When comparing the two models, the posterior mean of the coefficient from ONI_{MJJ} was slightly higher than that from ONI_{OND}, meaning a larger effect on disease onset. However, a larger SD was observed for the posterior distribution of the coefficient associated with ONI_{MJJ}, compared with the other trimesters, indicating a more variable effect of ONI_{MJJ}. The mixing parameters for the spatial effect (ϕ) of the three models were very similar (Table 4), which indicates that more than one-fourth of variability on the response variable is explained by the contribution of the spatial neighboring structure of municipalities, i.e. the risk of disease onset in one municipality is affected by neighboring municipalities that have already recorded SBR occurrences within the growing season.

Table 1: Mean, median, standard deviation (SD), 95% highest posterior density interval (HPDI), and posterior mode of model parameters and hyperparameters for the effect of Oceanic Ninõ Index (ONI) on time to onset of soybean rust in southern Brazil.

Trimester[a]	Parameters and hyperparameters[b]	Mean	Median	Mode	SD	$HPDI_{lower}$	$HPDI_{upper}$
JFM							
	β_0	-3.76	-3.76	-3.77	0.17	-4.09	-3.43
	ONI_{JFM}	0.08	0.08	0.08	0.16	-0.24	0.4
	λ	3.17	3	2.65	1.22	1.1	5.56
	τ	2.14×10^{9}	1.07×10^{8}	7.94×10^{6}	6.53×10^{10}	2.02×10^{3}	2.32×10^{11}
	ϕ	0.28	0.2	0.02	0.25	0	0.8
	δ	2.97×10^{1}	2.79×10^{1}	2.43×10^{1}	1.19×10^{1}	9.78	5.30×10^{1}
MJJ							
	β_0	-3.79	-3.79	-3.79	0.14	-4.07	-3.5
	ONI_{MJJ}	0.7	0.7	0.7	0.26	0.18	1.21
	λ	4.91	4.63	4.06	1.92	1.68	8.67
	τ	1.73×10^{9}	9.66×10^{7}	6.57×10^{6}	4.49×10^{10}	7.62×10^{3}	1.47×10^{11}
	ϕ	0.28	0.2	0.02	0.25	0	0.79
	δ	2.97×10^{1}	2.79×10^{1}	2.42×10^{1}	1.19×10^{1}	9.76	5.30×10^{1}
OND							
	β_0	-3.79	-3.79	-3.79	0.14	-4.06	-3.52
	ONI_{OND}	0.32	0.32	0.32	0.1	0.13	0.52
	λ	5.8	5.47	4.8	2.26	1.99	1.02×10^{1}
	τ	2.16×10^{9}	1.11×10^{8}	7.94×10^{6}	6.36×10^{10}	1.88×10^{3}	2.30×10^{11}
	ϕ	0.28	0.2	0.01	0.25	0	0.8
	δ	2.98×10^{1}	2.80×10^{1}	2.44×10^{1}	1.19×10^{1}	9.83	5.31×10^{1}

[a] *JFM* = January-February-March, MJJ = May-June-July, OND = October-November-December.
[b] β_0 = Model intercept; ONI_{JFM} = Oceanic Niño index for January-February-March; ONI_{MJJ} = Oceanic Niño index for May-June-July; ONI_{OND} = Oceanic Niño index for October-November-December; λ = Precision for seasons as random effects; τ= Precision for the spatial effect; ϕ = Mixing parameter for the spatial effect; δ= Precision for the baseline hazard function.

DISCUSSION

In this study, we explored the large database of 2,441 first reports of soybean rust in commercial fields at the municipality level within 17 growing seasons (2005 to 2021) in the three southernmost Brazilian states. Considerable variability in the timing of disease onset was

uncovered. Results from spatially structured Bayesian survival models indicated that the timing of disease onset was influenced by anomalies in the normal SST of the El Niño 3.4 region in the central Pacific before and during the crop season: MJJ and OND, respectively. The increase in ONI (e.g: warmer SST than normal) in these trimesters is expected to trigger the early onset of SBR epidemics in southern Brazil.

Results showed that soybean rust was detected in commercial fields as early as September, the beginning of the growing season in Brazil) or as late as mid-March (Fig. 1). The variation in detection timing may be due to several reasons, especially those linked to inoculum availability. For example, prior to the establishment of the soybean-free period in Brazil, it was common to detect the disease during the vegetative stages (Godoy et al. 2016). Crops grown during the off-season (winter) allowed the inoculum to build up and, consequently, spread and reach the crops grown in the main growing season (Godoy et al. 2016). Another reason is related to the early onset of rain and maintenance of an abnormally wet condition early in the growing season, which influences the soybean sowing schedule and also may boost inoculum production and infections early in the season.

The effect or association of ENSO with epidemic data (observation or simulation) was previously investigated for coffee leaf rust in Brazil (Hinnah et al. 2020), Fusarium head blight in wheat in the United States (Kriss et al. 2012), and wheat rust in China and in the United States (Scherm and Yang 1995). All these studies, although using different methodologies, highlighted the effects of El Niño (Hinna et al 2020) or La Niña (Kriss et al. 2012) in increasing the risk of the disease under study. Furthermore, Kriss et al. (2012) and Hinnah et al. (2020) also demonstrated the differential effects of the ENSO across different regions or states. This was more evident for coffee leaf rust, which showed that the differences in predicted infection rates between warm (El Niño) and cold (La Niña) phases were only significant in states further south of Brazil, the region in which we focused our study.

Previous works on the effects of ENSO-related variables on SBR were based on simulation of disease prediction/risk models using relatively long historical observations of weather data (Radons et al. 2021; Del Ponte et al. 2011; Fattori et al. 2021). For instance, Del Ponte et al. (2011) used a rainfall-based model (Del Ponte et al. 2006c) to predict final severity levels across 30 years (1979–2008) of rainfall observation in several locations across Rio Grande do Sul (RS), the southernmost state in the Southern Brazilian region. In another effort, also targeting the same region (RS), Radons et al. (2020) used an infection-based model (that uses air temperature and leaf wetness duration) to predict a “climate risk”, from which they derived the number of fungicide applications per cycle. Contrary to Del Ponte et al. (2006),

who found that SBR severity levels were generally increased during El Niño, and more strongly correlated depending on the location, the effect of the ENSO phases on the predicted number of applications per crop season across the locations studied was not clear (Radons et al. 2021). By linking a process-based soybean model and a rainfall-based SBR model adjusted to predict daily severity progress that penalized simulated yields, Fattori et al (2021) predicted increased severity levels during El Niño years. Consequently, higher yield losses were reported in Southern Brazil than in any other mega-production region of Brazil, agreeing with Del Ponte et al. (2011).

Using SBR-presence data, also reported by the consortium, Minchio et al. (2018) explored the effects of weather and ENSO-related variables on the final prevalence data reported for two states in Brazil: Mato Grosso (Mid-Western Region) and Paraná, both spanning 11 seasons (from 2004/05 to 2014/15 seasons). The SST values from August to December were included in the model as a predictor of the prevalence in Paraná State, suggesting that warmer SST (El Niño) should influence the regional spread of SBR. Our results are in general agreement with those. However, besides the differences in the modeling approach, those authors kept multiple occurrences for unique municipalities in the database, which inflates the number of occurrences in a single location. Furthermore, the absence of spatial effects (effects between municipalities) in their models may lead to misleading inferences (Besag 1974; Ojiambo and Kang 2013; Alves et al. 2021; Cendoya et al. 2020).

We found that the risk of SBR outbreaks was positively affected by ONI values during the OND and MJJ trimesters. No relationship between the risk of SBR outbreaks and the ONI values in JFM was observed. The effect of the ONI_{OND} was not unexpected because of its effect on the rainfall during the growing season (late spring and early summer) (Del Ponte et al. 2011). However, our model using ONI_{MJJ} provided a higher posterior mean effect of the anomaly in that period (compared to OND) on the risk of SBR onset. Possibly, the increase in the risk may be associated with an early onset of rainfall previously to the crop season, which boosts inoculum production on alternative hosts and voluntary soybean plants. This result is encouraging to allow an early warning of soybean rust at least three months prior to the start of the season. Similarly, Kriss et al. (2020) found that higher intensities of Fusarium head blight epidemics in wheat in Ohio, United States, would be expected one year after observing low ONI values, i.e., La Niña. Our results provide novel insights into the mechanisms and factors that drive the large-scale spread of soybean rust across Southern Brazil, which may be useful for seasonal outlook and decision-making in soybean rust management such as the timing of fungicide applications.

ACKNOWLEDGMENTS

The First author is thankful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES) for a scholarship. The second author is thankful to the CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for providing a research fellowship. We thank the leaders of the Consórcio Antiferrugem for providing the raw data used in this study.

DATA AVAILABILITY STATEMENT

The data and codes that support the findings of this study are openly available in the Open Science Framework at https://doi.org/10.17605/OSF.IO/2KHFV.

REFERENCES

Allen, T., Hollier, C., and Sikora, E. 2014. A Continuing Saga: Soybean Rust in the Continental United States, 2004 to 2013. Outlooks Pest Manag. 25:167–174.

Alves, K. S., Rothmann, L., and Del Ponte, E. 2021. Linking Climate Variables to Large-Scale Spatial Pattern and Risk of Citrus Huanglongbing: A Hierarchical Bayesian Modeling Approach. Phytopathology®. Available at: https://apsjournals.apsnet.org/doi/10.1094/PHYTO-05-21-0219-FI [Accessed September 3, 2021].

Alves, M. de C., de Carvalho, L. G., Pozza, E. A., Sanches, L., and Maia, J. C. de S. 2011. Ecological zoning of soybean rust, coffee rust and banana black sigatoka based on Brazilian climate changes. Procedia Environ. Sci. 6:35–49.

Anyamba, A., Chretien, J.-P., Britch, S. C., Soebiyanto, R. P., Small, J. L., Jepsen, R., et al. 2019. Global Disease Outbreaks Associated with the 2015–2016 El Niño Event. Sci. Rep. 9:1930.

Barro, J. P., Alves, K. S., Godoy, C. V., Dias, A. R., Forcelini, C. A., Utiamada, C. M., et al. 2021. Performance of dual and triple fungicide premixes for managing soybean rust across years and regions in Brazil: A meta-analysis. Plant Pathol. 70:1920–1935.

Beruski, G. C., Del Ponte, E. M., Pereira, André. B., Gleason, M. L., Câmara, G. M. S.,

Araújo Junior, I. P., et al. 2020. Performance and Profitability of Rain-Based Thresholds for Timing Fungicide Applications in Soybean Rust Control. Plant Dis. 104:2704–2712.

Besag, J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. J. R. Stat. Soc. Ser. B Methodol. 36:192–225.

Besag, J., York, J., and Mollié, A. 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 43:1–20.

Cai, W., McPhaden, M. J., Grimm, A. M., Rodrigues, R. R., Taschetto, A. S., Garreaud, R. D., et al. 2020. Climate impacts of the El Niño–Southern Oscillation on South America. Nat. Rev. Earth Environ. 1:215–231.

Cendoya, M., Martínez-Minaya, J., Dalmau, V., Ferrer, A., Saponari, M., Conesa, D., et al. 2020. Spatial Bayesian Modeling Applied to the Surveys of Xylella fastidiosa in Alicante (Spain) and Apulia (Italy). Front. Plant Sci. 11 Available at: https://www.frontiersin.org/articles/10.3389/fpls.2020.01204/full [Accessed August 24, 2020].

Cirino, P. H., Féres, J. G., Braga, M. J., and Reis, E. 2015. Assessing the Impacts of ENSO-related Weather Effects on the Brazilian Agriculture. Procedia Econ. Finance. 24:146–155.

Dalla Lana, F., Paul, P. A., Godoy, C. V., Utiamada, C. M., da Silva, L. H. C. P., Siqueri, F. V., et al. 2018. Meta-Analytic Modeling of the Decline in Performance of Fungicides for Managing Soybean Rust after a Decade of Use in Brazil. Plant Dis. 102:807–817.

Dalla Lana, F., Ziegelmann, P. K., Maia, A. de H. N., Godoy, C. V., and Del Ponte, E. M. 2015. Meta-Analysis of the Relationship Between Crop Yield and Soybean Rust Severity. Phytopathology. 105:307–315.

Del Ponte, E. M., Canteri, M. G., Reis, E. M., Yang, X. B., and Godoy, C. V. 2006a. Models and applications for risk assessment and prediction of Asian soybean rust epidemics. Fitopatol. Bras. 31:533–544.

Del Ponte, E. M., and Esker, P. D. 2008. Meteorological factors and Asian soybean rust epidemics: a systems approach and implications for risk assessment. Sci. Agric. 65:88–97.

Del Ponte, E. M., Godoy, C. V., Canteri, M. G., Reis, E. M., and Yang, X. B. 2006b. Models

and applications for risk assessment and prediction of Asian soybean rust epidemics. Fitopatol. Bras. 31:533–544.

Del Ponte, E. M., Godoy, C. V., Li, X., and Yang, X. B. 2006c. Predicting Severity of Asian Soybean Rust Epidemics with Empirical Rainfall Models. Phytopathology. 96:797–803.

Del Ponte, E. M., Maia, A. de H. N., dos Santos, T. V., Martins, E. J., and Baethgen, W. E. 2011. Early-season warning of soybean rust regional epidemics using El Niño Southern/Oscillation information. Int. J. Biometeorol. 55:575–583.

Fattori, I. M., Sentelhas, P. C., and Marin, F. R. 2021. Assessing the Impact of Climate Variability on Asian Rust Severity and Soybean Yields in Different Brazilian Mega-Regions. Int. J. Plant Prod. Available at: https://link.springer.com/epdf/10.1007/s42106-021-00169-x [Accessed October 21, 2021].

Godoy, C. V., Seixas, C. D. S., Soares, R. M., Marcelino-Guimarães, F. C., Meyer, M. C., and Costamilan, L. M. 2016. Asian soybean rust in Brazil: past, present, and future. Pesqui. Agropecuária Bras. 51:407–421.

Gómez-Rubio, V. 2020. *Bayesian inference with INLA*. Available at: http://becarioprecario.bitbucket.io/inla-gitbook/index.html [Accessed September 3, 2021].

Grimm, A. M. 2003. The El Niño Impact on the Summer Monsoon in Brazil: Regional Processes versus Remote Influences. J. Clim. 16:263–280.

Hinnah, F. D., Sentelhas, P. C., Gleason, M. L., Dixon, P. M., and Zhang, X. 2020. Assessing Biogeography of Coffee Rust Risk in Brazil as Affected by the El Niño Southern Oscillation. Plant Dis. 104:1013–1018.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. 2017. Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. J. Clim. 30:8179–8205.

Isard, S. A., Barnes, C. W., Hambleton, S., Ariatti, A., Russo, J. M., Tenuta, A., et al. 2011. Predicting Soybean Rust Incursions into the North American Continental Interior Using Crop Monitoring, Spore Trapping, and Aerobiological Modeling. Plant Dis. 95:1346–1357.

Koch, E., and Hoppe, H. H. 1987. Effect of Light on Uredospore Germination and Germ Tube Growth of Soybean Rust (Phakopsora pachyrhizi Syd.). J. Phytopathol. 119:64–74.

Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E., and Haines, A. 2003. El Niño and health. The Lancet. 362:1481–1489.

Kriss, A. B., Paul, P. A., and Madden, L. V. 2012. Variability in Fusarium Head Blight Epidemics in Relation to Global Climate Fluctuations as Represented by the El Niño-Southern Oscillation and Other Atmospheric Patterns. Phytopathology®. 102:55–64.

Lee, E. T., and Wang, J. W. 2003. *Statistical methods for survival data analysis*. 3rd ed. New York: J. Wiley.

Lindgren, F., and Rue, H. 2015. Bayesian Spatial Modelling with R-INLA. J. Stat. Softw. 63:1–25.

Machado, F. J., Barro, J. P., Godoy, C. V., Dias, A. R., Forcelini, C. A., Utiamada, C. M., et al. 2021. Is tank mixing site-specific premixes and multi-site fungicides effective and economic for managing soybean rust? a meta-analysis. Crop Prot. :105839.

Marchetti, M. A. 1976. The Effects of Temperature and Dew Period on Germination and Infection by Uredospores of Phakopsora pachyrhizi. Phytopathology. 66:461.

Minchio, C. A., Fantin, L. H., Caviglione, J. H., Braga, K., Silva, M. A. A. e, and Canteri, M. G. 2018. Predicting Asian Soybean Rust Epidemics Based on Off-Season Occurrence and El Niño Southern Oscillation Phenomenon in Paraná and Mato Grosso States, Brazil. J. Agric. Sci. 10:562.

NOAA (National Oceanic and Atmospheric Administration). 2022. Cold & Warm Episodes by Season. Available at: https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php [Accessed July 14, 2022].

Ojiambo, P. S., and Kang, E. L. 2013. Modeling Spatial Frailties in Survival Analysis of Cucurbit Downy Mildew Epidemics. Phytopathology. 103:216–227.

Pan, Z., Yang, X. B., Pivonia, S., Xue, L., Pasken, R., and Roads, J. 2006. Long-Term Prediction of Soybean Rust Entry into the Continental United States. Plant Dis. 90:840–846.

Radons, S. Z., Heldwein, A. B., Puhl, A. J., Nied, A. H., and da Silva, J. R. 2021. Climate risk

of Asian soybean rust occurrence in the state of Rio Grande do Sul, Brazil. Trop. Plant Pathol. 46:435–442.

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., and Chivian, E. 2001. Climate Change and Extreme Weather Events; Implications for Food Production, Plant Diseases, and Pests. Glob. Change Hum. Health. 2:90–104.

Rue, H., and Held, L. 2005. *Gaussian Markov random fields: theory and applications*. Boca Raton: Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71:319–392.

Scherm, H., and Yang, X. B. 1995. Interannual Variations in Wheat Rust Development in China and the United States in Relation to the El Nino/Southern Oscillation. Phytopathology. 85:970.

Sikora, E. J., Allen, T. W., Wise, K. A., Bergstrom, G., Bradley, C. A., Bond, J., et al. 2014. A Coordinated Effort to Manage Soybean Rust in North America: A Success Story in Soybean Disease Monitoring. Plant Dis. 98:864–875.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. 2017. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. Stat. Sci. 32:1–28.

Yorinori, J. T., Paiva, W. M., Frederick, R. D., Costamilan, L. M., Bertagnolli, P. F., Hartman, G. E., et al. 2005. Epidemics of Soybean Rust (*Phakopsora pachyrhizi*) in Brazil and Paraguay from 2001 to 2003. Plant Dis. 89:675–677.

CHAPTER 3: From reanalysis data to inference: a framework for linking environment to plant disease epidemics at the regional scale

This manuscript has been submitted for publication.
Formatting follows the journal requirements.

ABSTRACT

Traditional linear models can be too simplistic for capturing the myriad of interactions that occur among the triad of plant, pathogen and environment that results in plant disease epidemics. Tree-based machine learning (ML) algorithms are an attractive modeling solution because they automatically capture interactions, but work best when trained on a large input predictor matrix. In this study, multiple environmental and soil factors were collated from freely available gridded datasets and downscaled to better match the locations of snap bean fields evaluated for white mold across the central and western New York (NY) landscape. Functional data analysis of downscaled weather time series relative to planting was used to extract succinct summaries associated with disease occurrence. Summaries were input into a data matrix for training an ensemble tree model (XGBoost) fitted to prevalence of white mold. Environmental variables were ranked based on the contribution of their SHapley Additive exPlanations values to the fitted model. Nonlinear effects of weather and soil variables within the flowering period were detected. Most unexpectedly, air and soil temperatures at planting and in the weeks thereafter were associated with disease, which is not manifest until at least 40 days later. This study used a workflow of downscaled, gridded weather data, predictor selection, and ML model interpretation to enhance the understanding of environmental effects at a regional scale on one of the most prevalent and economically important diseases of snap beans. The proposed conceptual framework is broadly applicable to the regional prediction of other plant diseases.

Keywords: *Sclerotinia sclerotiorum*; white mold; weather; soil properties; data science.

INTRODUCTION

Environment is a major determinant of the occurrence and spatiotemporal dynamics of plant disease epidemics. Weather variables such as precipitation, temperature, relative humidity, solar radiation, wind speed and direction are well-known influences on many pathogen life-

cycle processes and the disease triangle (Shaner 1981; Huber and Gillespie 1992; Garrett et al. 2022), which consists of the interrelationships between the plant host, the pathogen, and environment. Soil chemical and physical properties also play an important role in regulating soil microbial diversity, survival or suppression of soilborne pathogens; and also plant nutrition, which may affect the susceptibility of the host plant to infection (Grewal et al. 1996; Noble and Coventry 2005; Bonanomi et al. 2010). Although the relationship of individual variables with diseases can be discerned (from controlled laboratory experiments), disentangling their contributions in the field where multiple environmental factors are involved can be challenging. This is mainly because environmental variables affect both the host and pathogen, influencing processes such as the latent and infectious periods of the pathogen, plant growth, and maturation, and ultimately disease intensity. Therefore, substantial research efforts have been invested to understand and quantify how epidemics in cultivated plant populations are shaped by the surrounding environment (Bourke 1970; Del Ponte and Esker 2008; Gent et al. 2013).

Typically, studies on the effect of weather on plant diseases have relied mostly on data gathered from established-network or researcher-provided on-site, ground-based weather stations. The data collected are used to empirically derive models linking weather-related variables to some measure of disease intensity (*e.g.*, final disease severity, the area under disease progress curves, infection rate, spore count, etc) (Del Ponte and Esker 2008; Newlands 2018; Shah et al. 2019b). However, although on-ground weather stations are generally considered the gold standard for plant disease research, the available network is often too sparse and of insufficient spatial resolution for the coverage of multiple farms or fields within a region; more so in low to middle-income countries, and in sparsely populated rural areas (Mistry et al. 2022; Auffhammer et al. 2013). Furthermore, adding new stations to the network is costly, and there are also potential issues of data recording gaps (either in space or time), equipment calibration, and data access. Climate reanalysis products are an alternative data source that has recently become more accessible to the research community. These data are generated using global or regional forecasting models in combination with data assimilation algorithms to estimate multiple weather parameters in the spatial and temporal dimensions. They provide a variety of agriculturally-relevant environmental variables, current and archived, and are usually freely accessible (Mistry et al. 2022). An example of a reanalysis product is the *ERA5-Land hourly data from 1950 to present'* (herein referred to as ERA5-land) dataset (Muñoz Sabater 2019), released by the European Centre for Medium-Range Weather Forecasts (ECMWF), and which provides reliable gridded data at a $0.1° \times 0.1°$ (9 km) spatial resolution.

Besides the challenges of sparsity and incomplete spatial coverage, the relationship between environment and disease intensity in agricultural fields is usually more complex than what simple linear models can capture (Cunniffe et al. 2015). This is because of the potentially high number of non-linear associations, variable interactions, and spatiotemporal associations that are present among environmental drivers and disease during the crop season. However, machine learning (ML) algorithms, such as ensemble tree models (*e.g.*, bagging, random forests, and gradient-boosted decision trees) and artificial neural networks, have rapidly gained popularity in the agricultural and broader ecological sciences for their capability to incorporate a large number of predictor variables (including those with missing data), nonlinear effects, and variable interactions (Humphries et al. 2018). Until recently, ML algorithms were largely viewed as black-box algorithms, for which interpretation of a predictor's influence on the response variable was practically impossible, unlike with traditional linear models. Yet, model interpretability is an important aspect not only to researchers but to stakeholders and decision-makers. Recent advances in ML model interpretation have removed that black-box label (Lundberg et al. 2020; Samek 2020). Tree-based ML algorithms have been used successfully for predicting plant disease epidemics in recent years while allowing an assessment of predictor contributions in the trained models (Shah et al. 2014, 2019a).

In this study, we propose a workflow of environmental data acquisition, data fusion, ML-based model building and interpretation for a more in-depth understanding of plant disease prevalence. Gridded sources of data from online databases (soil properties and weather data) were downscaled to the field level (Fig. 1) and fused with ground-based within-field disease observational data. The resulting data matrix was modeled using a tree-based ML algorithm with disease presence/absence as the response, and the fitted model interpreted with the goal of understanding the effects of weather and soil variables on the prevalence of disease at a regional scale. We additionally present an approach to predictor selection using functional data analysis of weather time series relative to crop planting to identify time-frames in which specific weather variables are associated with disease prevalence.

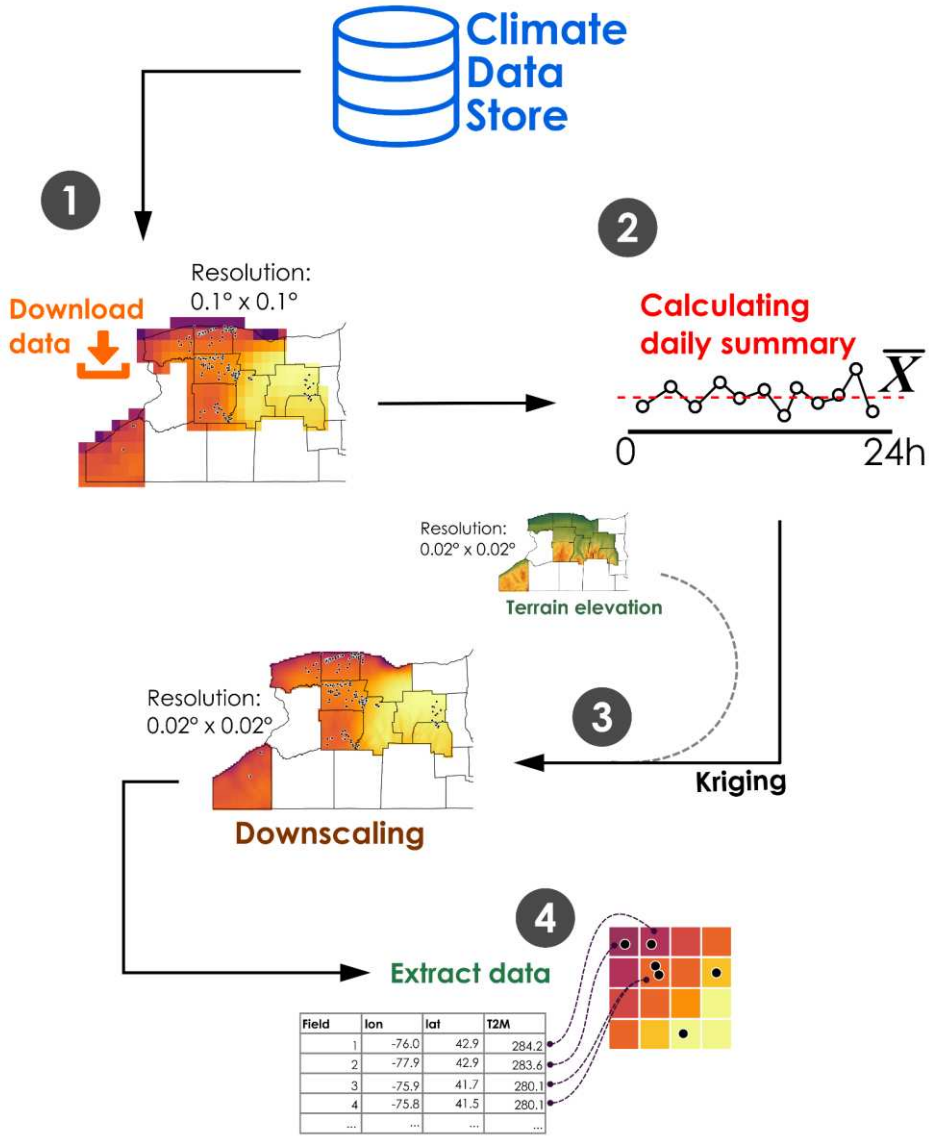

Fig. 1. Flow chart of steps to obtain and extract weather data for a field location. (1) The *ERA5-Land hourly data from 1950 to present'* dataset for the planting season was downloaded from the Climate Data Store; (2) The hourly data were then used to calculate the respective daily summaries; (3) Downscaling was performed using kriging, in which terrain elevation was used as a covariate; (4) The data for each field were extracted from the raster brick for the period 30 days before planting to 60 days after planting.

To demonstrate the proposed workflow, a case study was built using a three-year observational dataset of white mold prevalence in processing snap beans (*Phaseolus vulgaris* L.) in New York, USA (Fig. 2). White mold is caused by the soilborne fungus, *Sclerotinia sclerotiorum* (Lib.) de Bary (Willbur et al. 2018c). This disease has the unenviable reputation of being one of the most destructive diseases around the world, mainly due to the pathogen's broad host range which includes field (*e.g.* soybean) and specialty crops, the longevity of the primary inoculum source (sclerotia; which are hard, melanized, pebble-like structures) in the

soil, and considerable crop losses incurred (Willbur et al. 2018c; Derbyshire et al. 2022; Lehner et al. 2016). The sole infection pathway of *S. sclerotiorum* in leguminous crops is through the colonization of flower petals by ascospores. Ascospores are released from cup-like apothecia produced at the soil surface upon carpogenic germination of the sclerotia present in the soil. The dying petals serve as a nutrient source for the germinating ascospores which then infect other plant parts, such as pods and stems as the fungus develops (Saharan and Mehta 2008; Willbur et al. 2018c). As snap beans are determinate, in which flowers develop synchronously over a specific period, *S. sclerotiorum* has a restricted and short window in which to produce and disperse ascospores capable of infecting petals. Hence, the weather and soil conditions prior to and during flowering should play an important role in germination, ascospore dispersal, infection efficiency, and ultimately disease progress (Clarkson et al. 2004; Willbur et al. 2018a), thereby dictating disease prevalence at the regional scale.

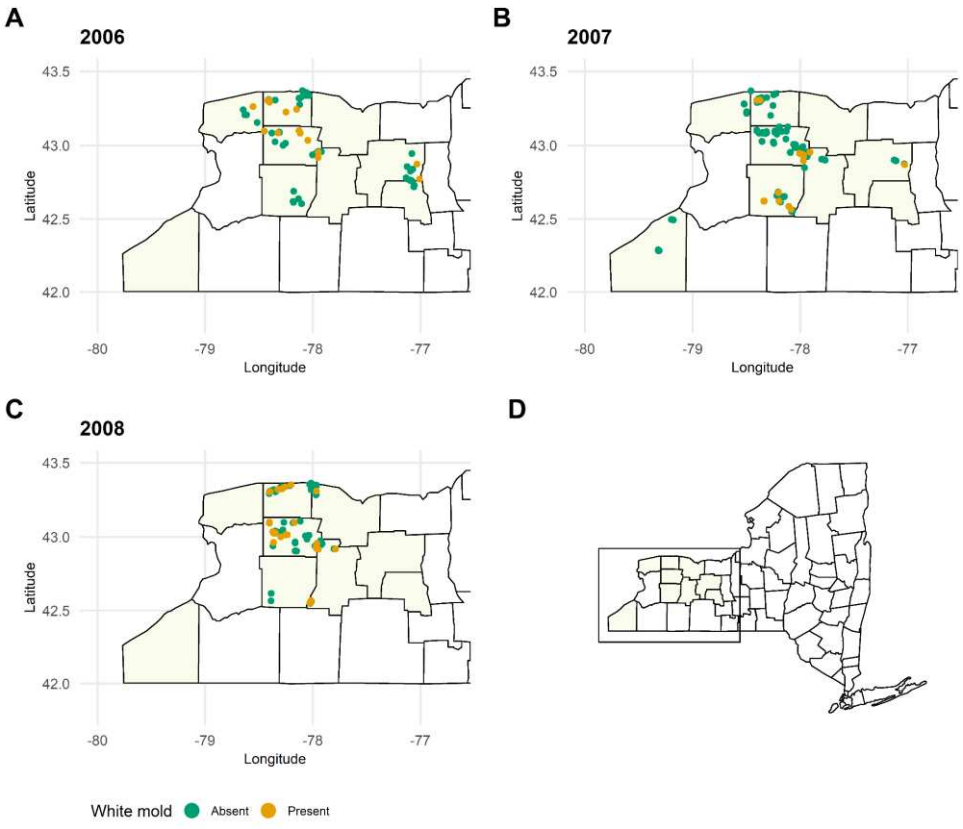

Fig. 2. Locations of commercial snap bean fields surveyed for white mold in counties across western and central New York, USA in (**A**) 2006, (**B**) 2007, and (**C**) 2008; (**D**) State map with counties divisions, in which the rectangle highlights the western and central region plotted in the previous panels.

RESULTS

Weather periods associated with disease presence

Flowering (bloom) occurred within 35 to 50 days after planting (dap) (*SI Appendix*, Fig. S1). Significant differences between the mean functional curves for the two disease classes (white mold prevalence: present or absent in a field) were observed during the bloom stage in all temperature-related variables (mean, minimum, and maximum air temperatures and dew point 2 m above the ground, as well as soil temperature within the upper 0 to 7 cm; Fig. 3A-E; see also *SI Appendix*, Fig. S2). Snap bean fields with white mold experienced cooler soil and air temperatures during flowering compared to fields in which white mold was absent. Significant differences between the mean functional curves for these variables were also observed before flowering, approximately 10 to 20 dap. However, the greater differences between the curves were observed from approximately 2 weeks prior to planting to 1 week after planting. For all temperature-related variables, there were differences of up to 1°C or higher between the mean functional curves (Fig. 3A-E), indicating that higher (soil and air) temperatures around the time of planting increase the chances of white mold in snap bean fields. In contrast to soil temperature, the mean functional curve for soil moisture in fields with white mold was consistently higher than the mean curve for fields without white mold the entire crop growing season, indicating that soil moisture is a crucial driver for white mold development in the field. Mean differences in moisture in the upper 7 cm of the soil varied from approximately 0.01 to 0.025 m^3/m^3. Higher differences were found within the flowering period (Fig. 3F). On the other hand, surface pressure tended to be lower across fields with white mold, compared to fields without the disease (Fig. 3G). The mean differences in surface pressure were almost constant across the growing season (around -250 Pa). Differences between the smoothed mean curves of relative humidity were only observed after flowering (Fig. 3H). Higher humidity after flowering could be expected to enhance white mold development in the field.

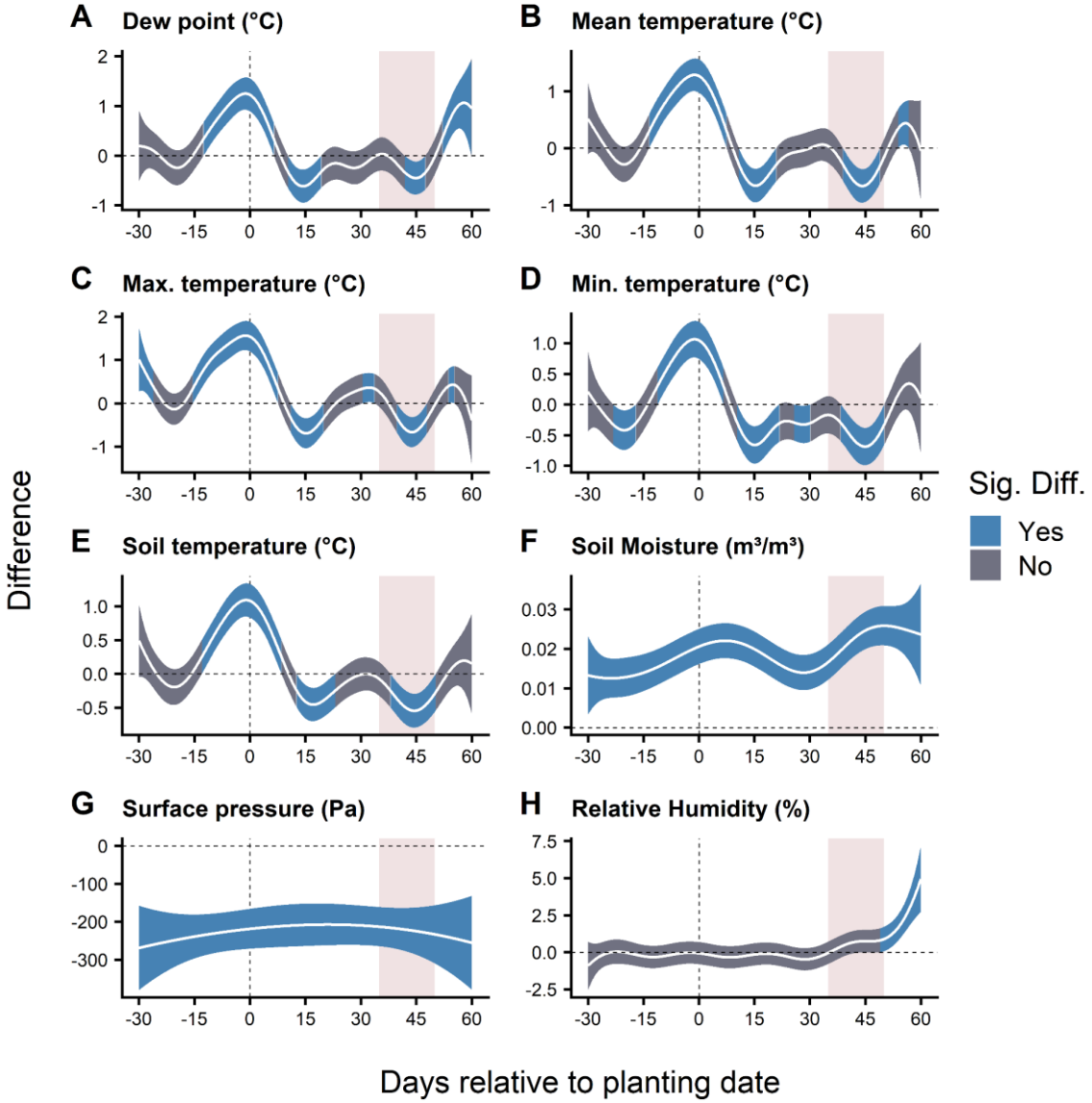

Fig. 3: Mean differences (solid white curves) and 95% confidence intervals (shaded ribbon) between functionally-represented weather time series for snap bean fields with and without white mold in New York, USA. The difference was estimated by subtracting the mean functional curve for fields without white mold from the mean curve for fields in which white mold was present. The rectangular shaded area between 35 and 50 days represents the snap bean bloom (flowering) period. Significant differences between the mean curves for the two classes of fields (Sig. Diff.) were interpreted to be present when confidence intervals did not include zero. (**A**) Daily mean dew point temperature at 2 meters above ground (2 m); (**B**) Daily mean temperature at 2 m; (**C**) Daily maximum temperature at 2 m; (**D**) Daily minimum temperature at 2 m; (**E**) Daily mean soil temperature within the upper 0 to 7 cm; (**F**) Daily mean soil moisture within the upper 0 to 7 cm; (**G**) Daily mean surface pressure; (**H**) Daily mean relative humidity at 2 m.

Model interpretation

An XGBoost was fitted, using soil and weather predictors, to the presence (and absence) of white mold in snap bean fields. The trained XGBoost model fit the data well, having an area under the receiver operating characteristic curve of 0.995. Model interpretation was via SHAP

values (Lundberg et al. 2020). Negative SHAP values are associated with a decrease in the risk of disease occurrence, while positive SHAP values are associated with an increase in the probability of white mold being present. SHAP values equal to or near zero are indicative of a predictor having no (or minimal) effect on the model's estimate of disease risk. Two approaches were applied to the analysis of the SHAP values: the SHAP summary plot, which ranks the predictor variables by their contributions to the model output; and SHAP dependence plots, which depict the SHAP values against the corresponding predictor value for each observation; they can be used to infer the effect of the predictor on the response variable (Lundberg et al. 2020).

Soil moisture (SM), soil temperature (ST), relative humidity (RH), and air temperature (T2M) in different time frames, and some soil properties featured among the top 10 most important predictors (Fig. 4A) because they had the highest absolute SHAP values (Fig. 4B). Soil moisture in the upper layers, throughout the entire growing period ($SM^{mean}_{[-15:50]}$), was the main contributor to the risk of white mold. T2M at planting and during the early development of the crop, as well as during flowering, was also an important white mold risk factor. Other contributing factors were soil organic matter content and ST during the early stages of crop development. RH during flowering and early pod development was also identified as an important contributor to white mold risk. Soil pH in water and bulk density likewise featured among the top 10 most influential variables associated with white mold risk.

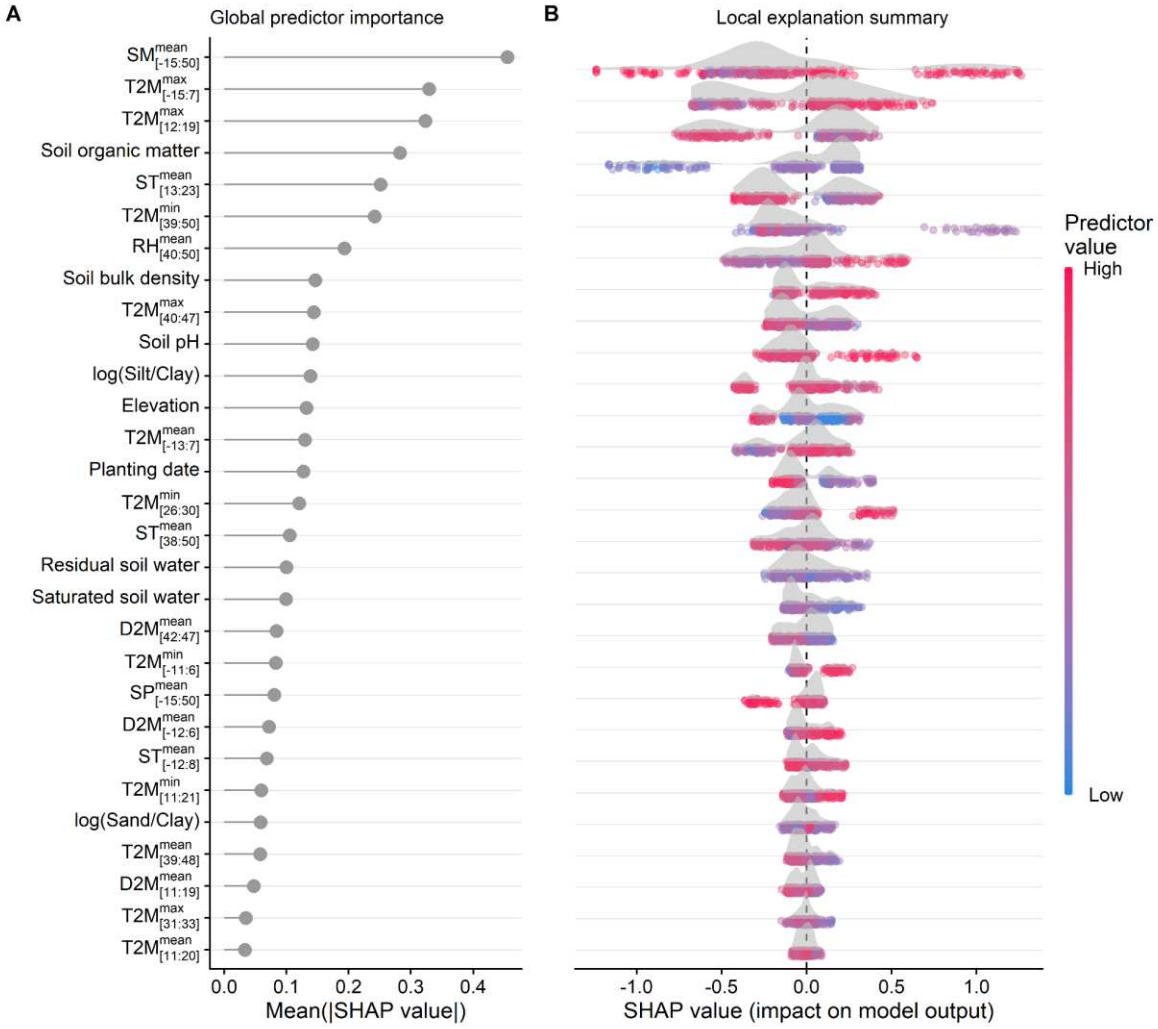

Fig. 4. SHapley Additive exPlanations (SHAP) summary plots of the 29 environmental predictors used in a trained XGBoost model for predicting white mold prevalence. (**A**) Mean absolute SHAP value for each of the environmental variables, a global measure of importance in predicting the outcome; (**B**) SHAP values for each observation for each environmental predictor. Higher SHAP values are associated with an increased probability of white mold presence in a field. SM = soil moisture within the upper 0 to 7 cm; T2M = air temperature 2 m above ground; ST = soil temperature within the upper 0 to 7 cm; RH = air relative humidity 2 m above ground; SP = surface pressure; D2M = dew point temperature 2 m above ground. For the weather-based variables, the superscript indicates the type of daily value (mean, minimum, or maximum) that was averaged over the period (in days) relative to planting indicated in the subscript (*e.g.*, [40:47] is the period from 40 to 47 days after planting).

SHAP dependence plots visualize the impact of each predictor value on increasing or decreasing disease risk (Fig. 5). We observed that dryer soils ($SM^{mean}_{[-15:50]}$ < 0.3 m^3/m^3) were associated with a decrease in disease risk, while $SM^{mean}_{[-15:50]}$ between 0.3 to 0.35 m^3/m^3 was associated with an increase in the probability of white mold presence (Fig. 5A). However, values higher than 0.35 m^3/m^3 appeared to be unfavorable to disease occurrence, even more so

than when $SM_{[-15:50]}^{mean}$ < 0.3 m^3/m^3. This indicates an optimal soil moisture range for white mold risk.

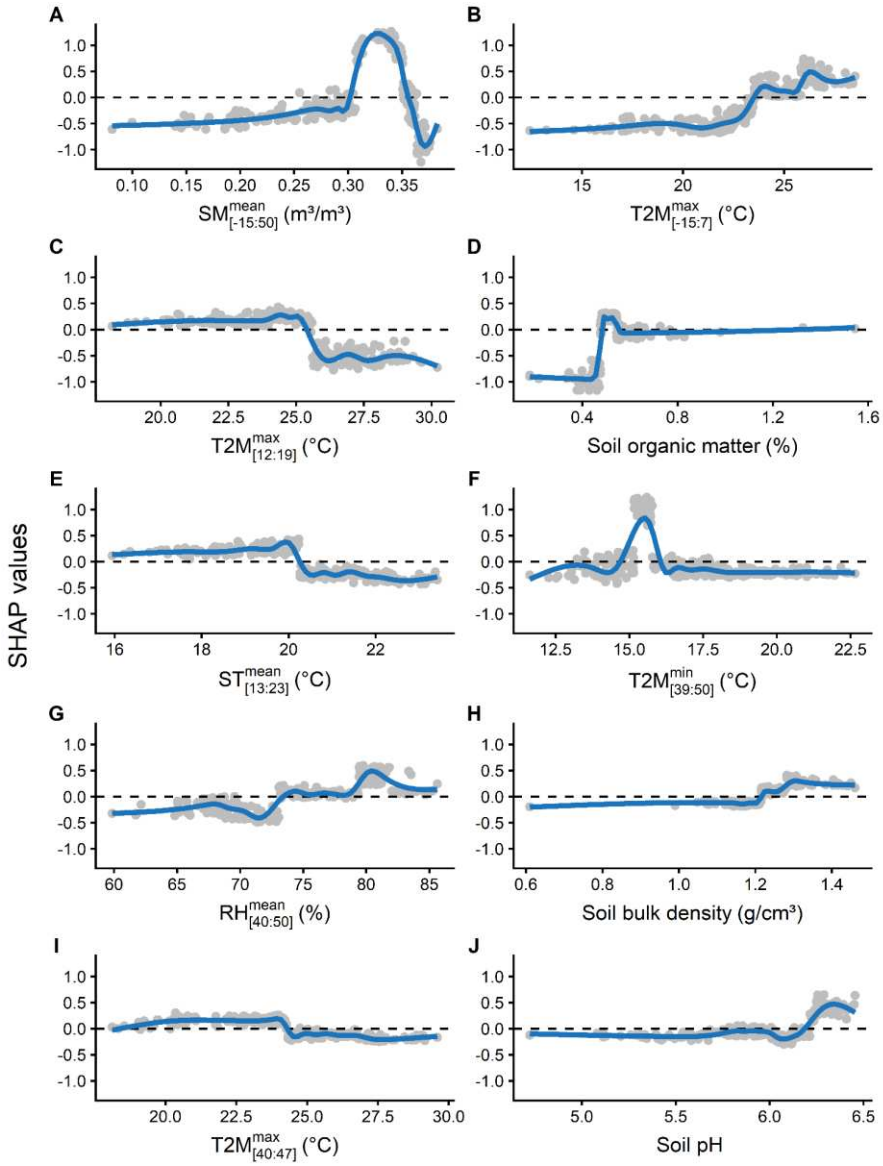

Fig. 5. SHapley Additive exPlanations (SHAP) dependence plots for the ten most important environmental variables in an XGBoost model predicting the presence of white mold in snap beans. Higher SHAP values are associated with an increased probability of white mold presence in the field. Solid lines represent cubic spline smooth curves fitted to the SHAP values. Points represent SHAP values estimated at each observed predictor value in the data (*x* axis). SM = soil moisture within the upper 0 to 7 cm; T2M = air temperature 2 m above ground; ST = Soil temperature within the upper 0 to 7 cm; RH = air relative humidity 2 m above ground; For the weather-based variables, the superscript indicates the type of daily value (mean, minimum, or maximum) that was averaged over the period (in days) relative to planting indicated in the subscript (*e.g.*, [40:47] is the period from 40 to 47 days after planting).

SHAP values increased nonlinearly with maximum temperatures around the planting date ($T2M^{max}_{[-15:7]}$). Cooler temperatures during this period (<23°C) were associated with a decrease in the risk of white mold. Values around ~23°C to ~26°C did not appear to have an effect on disease risk, with a slight increase in risk when it became warmer (>26°C) (Fig. 5B). Therefore, the difference between functional curves (Fig. 3C) was mostly due to lower $T2M^{max}_{[-15:7]}$ being associated with lower disease risk and not higher temperatures increasing the risk. Conversely, for $T2M^{max}_{[12:19]}$, which represents a week-long window about two weeks after planting, SHAP values were very close to zero until around 25°C, when they became negative (Fig. 5C), suggesting that daily maximum temperatures above 25°C in this 7-day period were detrimental to the germination of *S. sclerotiorum* sclerotia and subsequent disease risk, while values below 25°C have a small effect on the probability of white mold occurrence. The same pattern was observed for $ST^{mean}_{[13:23]}$ and $T2M^{max}_{[40:47]}$, but with ~20°C and ~24°C as thresholds (Fig. 5E and Fig. 5I).

Soil organic matter content lower than 0.5% was associated with a substantial decrease in disease risk, but values above 0.5% had no effect on increasing nor decreasing risk (Fig. 5D). Minimum temperatures within the flowering period ($T2M^{min}_{[39:50]}$) were more favorable to disease when between approximately 14°C to 16°C (Fig. 5F). Values outside of this range tended to have a negligible effect on disease risk. Nonlinear increases in SHAP values were also observed with RH increases in the flowering period (Fig. 5G). RH values below 75% were associated with a decrease in disease risk, whereas values above 80% appeared to favor disease development. Soil bulk density was associated with disease occurrence when higher than 1.2 g/m^3. Values below 1.2 g/m^3 had a small effect on decreasing risk (Fig. 5H). Soil pH was associated with white mold presence as it increased towards neutrality while being slightly unfavorable disease occurrence in acidic soils (Fig. 5J). SHAP dependence plots of the remaining predictor variables can be inspected in the *SI Appendix*, Fig S4.

DISCUSSION

A workflow for the integration of environmental data and field-level observational data on plant disease occurrence

We proposed a workflow for the fusion of environmental data downscaled to the same spatial extent at which a plant disease survey assessment would be made (presence or absence of disease in individual fields at the regional level). Our approach is based on retrieving freely-

available online gridded weather and soil data. Functional data analysis was used to identify periods within the crop season in which weather variables were highly associated with disease occurrence in a field. The data from these time frames were summarized and fused with soil properties, elevation, and planting date in a single matrix that was used to train an XGboost model for predicting white mold presence in snap bean fields in NY, USA.

Plant disease epidemiologists have long recognized the triad of host, pathogen, and environment, but have been hampered by the lack of high-quality environmental data at the spatial scale of their field studies. Access to environmental, agronomic, and weather data at finer spatial scales than the on-ground weather station networks has typically been out of reach of the applied agricultural researcher or has been proprietary (Magarey et al. 2001), but the recent increase in available API scripts is breaking down the barriers (Sparks 2018; Kusch and Davy 2022). Plant disease epidemiological studies making use of downscaled gridded data, from one or multiple sources, are infrequent (Seem 2004; Salam et al. 2011; Ouma et al. 2016). Our study demonstrated a comprehensive approach to accessing and processing online gridded data products and their application to the study of plant disease on a regional scale. The approach is expandable, both in terms of incorporating new data products and broadening the spatial extent beyond a few counties. Gridded and reanalysis data products in conjunction with downscaling now enable plant disease researchers and epidemiologists to obtain environmental data at spatial and temporal resolutions that will permit studies involving a higher number of fields or over larger geographical regions than traditionally possible, especially in areas with few on-ground stations. Nonetheless, it is important to realize that reanalysis data are always associated with some level of uncertainty (12), as is any weather data product, and this should be kept in mind when developing models based on such data (Mourtzinis et al. 2017).

A key feature of the workflow was the integration of downscaling with extended weather time series data (over the entire cropping season). Although entire weather time series could be used as a functional predictor of white mold presence (Shah et al. 2019b), we instead used FDA as a precursor to obtaining succinct weather variable summaries, as not all parts of a time series may be associated with a disease outcome; signals may be stronger or more relevant in some parts of a time series than in other sections. Plant disease epidemiologists have traditionally mined weather time series using some form of a window-pane algorithm (Coakley et al. 1988), which stratifies the time series data into discrete, fixed-time periods or windows. Summaries over these windows are then used as predictors (De Wolf et al. 2003; Kriss et al. 2010; Shah et al. 2013; Xu et al. 2013; Hill et al. 2019). However, building a set of candidate predictors using window-pane methodology inevitably generates a high-dimensional matrix of

potentially highly-correlated predictors (Shah et al. 2019b, 2018). The number of predictors also depends on the arbitrarily-chosen length of the window. We believe FDA to be a more principled approach to identifying periods that contain a signal associated with the target response. Consequently, fewer but potentially more relevant predictors are selected for input into model training.

Environmental variables associated with white mold occurrence

Having trained an XGBoost model with the environmental and soil variables, we next turned to model interpretation. Machine learning model interpretation can be done at the global (mean effect of a predictor over all observations) or local levels (contribution of each predictor to individual observations). Several different model-agnostic methods have been proposed for ML interpretation (Lundberg and Lee 2017; Hall and Gill 2018; Molnar 2022). Our study focused on SHAP values for interpreting predictor effects (Lundberg et al. 2020), via SHAP summary plots and SHAP dependence plots.

Soil moisture was the most important variable for predicting the occurrence of white mold in snap bean fields. The functional data analysis showed that this variable was important throughout the entire cropping season. The majority of fields were not irrigated (95.7% of fields), and therefore water availability was contingent on precipitation and soil water retention. Soil moisture is an important factor for apothecial formation (Clarkson et al. 2004; Abawi and Grogan 1975; Foster et al. 2011). However, previous studies have typically identified only the 1-2 weeks before flowering as being relevant with respect to the effects of soil moisture on white mold risk (Hunter et al. 1984). Relative humidity and air temperature during flowering also featured as important predictors, with elevated white mold risk especially when the minimum temperature was 14°C to 16°C and with increasing relative humidity. Earlier studies have also reported the general importance of temperature and soil moisture for *S. sclerotiorum* apothecial development in soybean (Willbur et al. 2018a, 2018b), carrot (Foster et al. 2011), and lettuce (Clarkson et al. 2004).

An intriguing result was that temperature at and soon after planting was associated with white mold risk. The probability of white mold occurrence increased with higher daily temperatures around planting, yet conversely, lower air (and soil) temperatures just two or three weeks after planting also were associated with a higher risk of white mold. It is difficult to hypothesize a plausible biological reason for the observation, assuming it is not a data artifact (which only additional observations can clarify). One possible explanation (among others) is a

host-inducing sclerotial dormancy-breaking or activation trigger (e.g., by seedling or root exudates) moderated by temperature. To date, no study has demonstrated such an interaction between *S. sclerotiorum* and snap beans. However, in sunflowers, the germination of sclerotia is likely triggered by seedling exudates (Rimmer and Menzies 1983).

Soil organic matter content was the most influential among the primary soil variables associated with white mold presence. Organic matter content below 0.5% was associated with a substantial decrease in disease risk. This result concords with those described between *S. sclerotiorum* and soil organic matter in another study with beans, where a higher number of *S. sclerotiorum* apothecia was observed in soils with a higher fraction of organic matter (Ferraz et al. 1999). Bulk density has a great effect on water infiltration and consequently on a soil's water retention capacity, mostly at superficial layers, which coincides with the location of the most epidemiologically-relevant sclerotia (Hunter et al. 1984; Ferraz et al. 1999). The results also suggest that sclerotia may do better when soil pH is closer to neutral. The effect of soil pH on *S. sclerotiorum* biology has not been thoroughly investigated, but a few studies have mentioned a possible interaction between pH and fungal development (Merriman 1976; Adams and Ayers 1979).

In conclusion, this study makes use of gridded environmental data coupled with downscaling, predictor selection, and ML model interpretation to disentangle the associations of environmental variables (soil and weather) with the regional occurrence of a plant disease. A corollary was a rethinking of commonly held perspectives on the interaction between *S. sclerotiorum* and one of its hosts, snap bean, suggesting new directions of research. The approach presented in this article can be used as a conceptual framework for revisiting paradigms of plant disease risk at the landscape level.

MATERIAL AND METHODS

Data on white mold prevalence

The prevalence (presence or absence) of white mold was assessed in commercial snap bean fields in New York (356 fields total) during 2006 (88 fields), 2007 (144 fields), and 2008 (124 fields) across western and central NY, USA (Fig. 2). *S. sclerotiorum* is known to be widespread in agricultural fields in this region (Abawi and Grogan 1975; Shah et al. 2019a). A snap bean field was considered positive for white mold if the disease was visually observed on any part (typically stems and pods) of a sampled plant, where at least 50 plants were observed per field.

Among all fields, 20% were classified as having white mold. Details surrounding the collection of the field-based observational data were presented earlier (Shah et al. 2019a).

Weather data

Reanalysis weather data were downloaded from the '*ERA5-Land hourly data from 1950 to present'* database (Muñoz Sabater 2019) via the Climate Data Store (CDS). Data were downloaded for April 1 to October 31 in each of the years 2006, 2007, and 2008. The variables retrieved are listed in *SI Appendix*, Table S1. As described in the ERA-Interim documentation (ECMWF 2007), the daily average of the air relative humidity (RH^{mean}; %) can be derived from the ratio between saturation vapor pressure calculated for the daily average temperature at 2 m above ground ($T2M^{mean}$, in Kelvin) and daily average dew point temperature at 2 m above ground ($D2M^{mean}$, in Kelvin).

$$RH^{mean} = \frac{es(D2M^{mean})}{es(T2M^{mean})} 100 \quad (1)$$

where $es(\cdot)$ represents the saturation vapor pressure function (Equation 2):

$$es(T) = \alpha_1 \, exp\left[\alpha_3 \left(\frac{T - t_0}{T - \alpha_4}\right)\right] \quad (2)$$

where, T is temperature (in Kelvin), $\alpha_1 = 611.21$, $\alpha_3 = 17.502$, $\alpha_4 = 32.19$ and $t_0 = 273.16$ K.

ERA5-Land weather data are available at a 0.1° (9 km) global spatial resolution, which were still too coarse relative to the spatial separation between the sampled snap bean fields and the topography of central and western NY state. To obtain data at a higher spatial resolution, weather data were downscaled to a 0.02° (~1.57 km) spatial resolution using kriging with elevation as a covariate in the model for all variables except soil moisture (in the upper 7 cm). In the case of soil moisture, the downscaled soil temperature was used as the covariate in the kriging model. The kriging process was performed in R version 4.2 (R Core Team 2022) using the 'krigR' package (Kusch and Davy 2022). A flow chart of the weather data processing is displayed in Fig. 1. Maps of downscaled results were examined visually for consistency and rationality based on what may be expected for central and western NYk, USA given the region's topography and proximity to the Great Lakes Erie and Ontario.

Soils data

Soil properties data were obtained from the POLARIS database (Chaney et al. 2019), which provides probabilistic estimates of soils data at a 30-m resolution within the contiguous US. We wanted to capture variables related to soil chemistry (pH and organic matter), soil texture (clay, silt, and sand content), soil density, and water retention (residual and saturated water content), as these may affect sclerotial survival and germination. A full detailed list of soil variables is available in *SI Appendix*, Table S1, and the distribution of each variable in western and central NY is depicted in *SI Appendix*, Fig. S3. As clay, silt, and sand are compositional data by nature (i.e., the three proportions must sum to 1, and any part can be calculated by knowing the other two parts), log-transformations of the ratios between sand and clay [log(sand/clay)] and silt and clay [log(silt/clay)] were used as predictors in the ML model instead of the original sand, silt, and clay values (Greenacre 2019). Seven soils-related predictor variables were derived. The data were downloaded using the R package 'XPolaris' (Rosso et al. 2021).

Functional data analysis

Daily weather time series, from 30 days before planting to the final field sampling date [50 to 77 days after planting (dap)], were estimated for each snap bean field after downscaling the ERA5-Land data by kriging (Fig. 5; see also *SI Appendix*, Table S1). Function-on-scalar regression (Morris 2015) was done with the objective of determining whether there were differences between the weather time series associated with snap bean fields with and without white mold. Functional data analysis was performed using generalized additive mixed-effects models, in which each field-year combination was modeled with random slope and intercept. Time-series data were represented by 4^{th} order penalized B-splines smoothed curves and with 17 knots. Model fitting was performed using the R package, 'mgcv' (Wood 2017). Visualization and comparison of the smoothed curves was performed using the R package 'tidymv' (Coretta 2022).

For predictor variable creation (for input into the XGBoost algorithm), weather time series were first subset to the period from 15 days before planting to 50 days after, which covers the end of the bloom and excludes the harvest period. Then, periods in which there were no significant differences between the mean functional curves were filtered out. That is, only data from periods with significant differences between the mean functional curves were retained (Fig. 2), the exception being for relative humidity, in which data from 40 to 50 dap were retained. Weather data within these periods were averaged and used as predictors in training the ML model. That is, the variables represented the average of a daily weather parameter

summary (minimum, maximum or mean) over the number of days in the window. We use the notation $W^{m}_{[a:b]}$ to describe the variable, where W is the weather parameter ($e.g.$, T2M), m is the type of daily summary (minimum, maximum or mean), and $[a:b]$ is the window in days (relative to planting) over which the daily summaries were averaged. Therefore for example, $T2M^{max}_{[5:10]}$ would represent the average of the daily maximum values of T2M over the window from five to 10 days after planting.

Model fitting and interpretation

We used the XGBoost ML algorithm to classify fields with and without white mold on 29 environmental variables. Twenty predictor variables were weather-related (formed from the results of the functional data analysis described above), seven were soils-related and two (field elevation and planting date) were taken from the original observational dataset (the full list of predictor variables is in *SI Appendix*, Table S1). A total of 60 combinations of randomly generated hyperparameters (i.e., a grid search space) were tested to find the model with the lowest mean logarithmic loss in five-fold cross-validation. The hyperparameters control tree size, sample sizes, and learning rates used for model training. The full list of finalized hyperparameter values used in the best model and their respective descriptions are in *SI Appendix*, Table S2. Tree depth was restricted to no more than three to avoid overfitting. As our study focused on model interpretation as opposed to prediction, the XGBoost model was trained using all 356 observations. The area under the receiver operating characteristic curve was used as a measure of model fit. The model fitting procedure was performed within the 'tidymodels' package framework (Kuhn and Wickham 2020). The fitted model was interpreted by estimating the SHapley Additive exPlanations (SHAP) associated with the variables for each of the observations, using the R package 'SHAPforxgboost' (Liu and Just 2021).

Data availability and reproducibility

Data and code supporting the findings of this study are available in an Open Science Framework repository (https://doi.org/10.17605/OSF.IO/V53PY).

ACKNOWLEDGMENTS

KSA is thankful to *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil* (CAPES) for a scholarship. EMD is thankful to the *Conselho Nacional de Desenvolvimento Científico e Tecnológico-Brasil* (CNPQ) for providing their research fellowship. This research

was also supported by the United States Department of Agriculture National Institute of Food and Agriculture Hatch project NYG-6253320, managed by Cornell AgriTech, Cornell University, Geneva, NY.

REFERENCES

Abawi, G. S., and Grogan, R. G. 1975. Source of primary inoculum and effects of temperature and moisture on infection of beans by *Whetzelinia sclerotiorum*. Phytopathology. 65:300.

Adams, P., and Ayers, W. 1979. Ecology of *Sclerotinia* species. Phytopathology. 69:896–899.

Auffhammer, M., Hsiang, S. M., Schlenker, W., and Sobel, A. 2013. Using weather data and climate model output in economic analyses of climate change. Rev. Environ. Econ. Policy. 7:181–198.

Bonanomi, G., Antignani, V., Capodilupo, M., and Scala, F. 2010. Identifying the characteristics of organic soil amendments that suppress soilborne plant diseases. Soil Biol. Biochem. 42:136–144.

Bourke, P. M. A. 1970. Use of weather information in the prediction of plant disease epiphytotics. Annu. Rev. Phytopathol. 8:345–370.

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., et al. 2019. POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. Water Resour. Res. 55:2916–2938.

Clarkson, J. P., Phelps, K., Whipps, J. M., Young, C. S., Smith, J. A., and Watling, M. 2004. Forecasting sclerotinia disease on lettuce: toward developing a prediction model for carpogenic germination of sclerotia. Phytopathology®. 94:268–279.

Coakley, S. M., McDaniel, L. R., and Line, R. F. 1988. Quantifying how climatic factors affect variation in plant disease severity: A general method using a new way to analyze meteorological data. Clim. Change. 12:57–75.

Coretta, S. 2022. *tidymv: tidy model visualisation for generalised additive models*. Available at: https://CRAN.R-project.org/package=tidymv.

Cunniffe, N. J., Koskella, B., E. Metcalf, C. J., Parnell, S., Gottwald, T. R., and Gilligan, C. A. 2015. Thirteen challenges in modelling plant diseases. Epidemics. 10:6–10.

De Wolf, E. D., Madden, L. V., and Lipps, P. E. 2003. Risk assessment models for wheat fusarium head blight epidemics based on within-season weather data. Phytopathology. 93:428–435.

Del Ponte, E. M., and Esker, P. D. 2008. Meteorological factors and Asian soybean rust epidemics: a systems approach and implications for risk assessment. Sci. Agric. 65:88–97.

Derbyshire, M. C., Newman, T. E., Khentry, Y., and Owolabi Taiwo, A. 2022. The evolutionary and molecular features of the broad-host-range plant pathogen *Sclerotinia sclerotiorum*. Mol. Plant Pathol. :mpp.13221.

ECMWF. 2007. IFS documentation CY31R1 - Part IV: Physical processes. In *IFS Documentation CY31R1*, IFS Documentation, ECMWF. Available at: https://www.ecmwf.int/node/9221.

Ferraz, L. C. L., Café Filho, A. C., Nasser, L. C. B., and Azevedo, J. 1999. Effects of soil moisture, organic matter and grass mulching on the carpogenic germination of sclerotia and infection of bean by *Sclerotinia sclerotiorum*. Plant Pathol. 48:77–82.

Foster, A. J., Kora, C., McDonald, M. R., and Boland, G. J. 2011. Development and validation of a disease forecast model for Sclerotinia rot of carrot. Can. J. Plant Pathol. 33:187–201.

Garrett, K. A., Bebber, D. P., Etherton, B. A., Gold, K. M., Sulá, A. I. P., and Selvaraj, M. G. 2022. Climate change effects on pathogen emergence: artificial intelligence to translate big data for mitigation. Annu. Rev. Phytopathol. 60 Available at: https://doi.org/10.1146/annurev-phyto-021021-042636 [Accessed June 14, 2022].

Gent, D. H., Mahaffee, W. F., McRoberts, N., and Pfender, W. F. 2013. The use and role of predictive systems in disease management. Annu. Rev. Phytopathol. 51:267–289.

Greenacre, M. J. 2019. *Compositional data analysis in practice*. Boca Raton: CRC Press, Taylor and Francis Group.

Grewal, H. S., Graham, R. D., and Rengel, Z. 1996. Genotypic variation in zinc efficiency and resistance to crown rot disease (*Fusarium graminearum* Schw. Group 1) in wheat. Plant Soil. 186:219–226.

Hall, P., and Gill, N. 2018. *Introduction to machine learning interpretability*. Place of publication not identified: O'Reilly Media, Inc. Available at:

https://proquest.safaribooksonline.com/9781492033158 [Accessed July 21, 2022].

Hill, G. N., Beresford, R. M., and Evans, K. J. 2019. Automated analysis of aggregated datasets to identify climatic predictors of botrytis bunch rot in wine grapes. Phytopathology®. 109:84–95.

Huber, L., and Gillespie, T. J. 1992. Modeling leaf wetness in relation to plant disease epidemiology. Annu. Rev. Phytopathol. 30:553–577.

Humphries, G., Magness, D. R., and Huettmann, F., eds. 2018. *Machine learning for ecology and sustainable natural resource management*. Cham: Springer International Publishing. Available at: http://link.springer.com/10.1007/978-3-319-96978-7 [Accessed July 21, 2022].

Hunter, J. E., Pearson, R. C., Seem, R. C., Smith, C. A., and Palumbo, D. R. 1984. Relationship between soil moisture and occurrence of *Sclerotinia sclerotiorum* and white mold disease on snap beans. Prot. Ecol. 7:269–280.

Kriss, A. B., Paul, P. A., and Madden, L. V. 2010. Relationship between yearly fluctuations in fusarium head blight intensity and environmental variables: a window-pane analysis. Phytopathology. 100:784–797.

Kuhn, M., and Wickham, H. 2020. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* Available at: https://www.tidymodels.org.

Kusch, E., and Davy, R. 2022. KrigR—a tool for downloading and statistically downscaling climate reanalysis data. Environ. Res. Lett. 17:024005.

Lehner, M. S., Pethybridge, S. J., Meyer, M. C., and Del Ponte, E. M. 2016. Meta-analytic modelling of the incidence-yield and incidence-sclerotial production relationships in soybean white mould epidemics. Plant Pathol. 66:460–468.

Liu, Y., and Just, A. 2021. *SHAPforxgboost: SHAP plots for "XGBoost."* Available at: https://CRAN.R-project.org/package=SHAPforxgboost.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2:56–67.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In

Advances in Neural Information Processing Systems, Curran Associates, Inc. Available at: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [Accessed June 21, 2022].

Magarey, R. D., Seem, R. C., Russo, J. M., Zack, J. W., Waight, K. T., Travis, J. W., et al. 2001. Site-specific weather information without on-site sensors. Plant Dis. 85:1216–1226.

Merriman, P. R. 1976. Survival of sclerotia of *Sclerotinia sclerotiorum* in soil. Soil Biol. Biochem. 8:385–389.

Mistry, M. N., Schneider, R., Masselot, P., Royé, D., Armstrong, B., Kyselý, J., et al. 2022. Comparison of weather station and climate reanalysis data for modelling temperature-related mortality. Sci. Rep. 12:5178.

Molnar, C. 2022. *Interpretable machine learning: a guide for making black box models explainable*.

Morris, J. S. 2015. Functional regression. Annu. Rev. Stat. Its Appl. 2:321–359.

Mourtzinis, S., Rattalino Edreira, J. I., Conley, S. P., and Grassini, P. 2017. From grid to field: Assessing quality of gridded weather data for agricultural applications. Eur. J. Agron. 82:163–172.

Muñoz Sabater, J. 2019. ERA5-Land hourly data from 2001 to present. Copernic. Clim. Change Serv. C3S Clim. Data Store CDS. Available at: https://doi.org/10.24381/cds.e2161bac [Accessed February 25, 2022].

Newlands, N. K. 2018. Model-based forecasting of agricultural crop disease risk at the regional scale, integrating airborne inoculum, environmental, and satellite-based monitoring data. Front. Environ. Sci. 6 Available at: https://www.frontiersin.org/article/10.3389/fenvs.2018.00063 [Accessed February 23, 2022].

Noble, R., and Coventry, E. 2005. Suppression of soil-borne plant diseases with composts: A review. Biocontrol Sci. Technol. 15:3–20.

Ouma, P. O., Odera, P. A., and Mukundi, J. B. 2016. Spatial Modelling of Weather Variables for Plant Disease Applications in Mwea Region. J. Geosci. Environ. Prot. 4:127–136.

R Core Team. 2022. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rimmer, S. R., and Menzies, J. G. 1983. Influence of host seedling exudates on germination of sclerotia of *Sclerotinia sclerotiorum* [rape and sunflower disease]. In *Proceedings of the 6th International Rapeseed Conference*, Paris, p. 951–956. Available at: https://www.gcirc.org/fileadmin/documents/Proceedings/IRC1983vol1/83Diseases/CO83%20-%20page%20951.pdf [Accessed June 22, 2022].

Rosso, L. H. M., Reis, A. F. de B., Correndo, A. A., and Ciampitti, I. A. 2021. XPolaris: an R-package to retrieve United States soil data at 30-meter resolution. BMC Res. Notes. 14:327.

Saharan, G. S., and Mehta, N. 2008. *Sclerotinia diseases of crop plants: biology, ecology and disease management*. Dordrecht: Springer.

Salam, M. U., MacLeod, W. J., Salam, K. P., Maling, T., and Barbetti, M. J. 2011. Impact of climate change in relation to ascochyta blight on field pea in Western Australia. Australas. Plant Pathol. 40:397.

Samek, W. 2020. Learning with explainable trees. Nat. Mach. Intell. 2:16–17.

Seem, R. C. 2004. Forecasting plant disease in a changing climate: a question of scale. Can. J. Plant Pathol. 26:274–283.

Shah, D. A., De Wolf, E. D., Paul, P. A., and Madden, L. V. 2018. Functional data analysis of weather variables linked to fusarium head blight epidemics in the United States. Phytopathology. 109:96–110.

Shah, D. A., De Wolf, E. D., Paul, P. A., and Madden, L. V. 2014. Predicting fusarium head blight epidemics with boosted regression trees. Phytopathology®. 104:702–714.

Shah, D. A., Dillard, H. R., and Pethybridge, S. J. 2019a. Identification of factors associated with white mould in snap bean using tree-based methods. Plant Pathol. 68:1694–1705.

Shah, D. A., Molineros, J. E., Paul, P. A., Willyerd, K. T., Madden, L. V., and De Wolf, E. D. 2013. Predicting fusarium head blight epidemics with weather-driven pre- and post-anthesis logistic regression models. Phytopathology. 103:906–919.

Shah, D. A., Paul, P. A., De Wolf, E. D., and Madden, L. V. 2019b. Predicting plant disease

epidemics from functionally represented weather series. Philos. Trans. R. Soc. B Biol. Sci. 374:20180273.

Shaner, G. 1981. Effect of environment on fungal leaf blights of small grains. Annu. Rev. Phytopathol. 19:273–296.

Sparks, A. H. 2018. nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. J. Open Source Softw. 3:1035.

Willbur, J. F., Fall, M. L., Bloomingdale, C., Byrne, A. M., Chapman, S. A., Isard, S. A., et al. 2018a. Weather-based models for assessing the risk of *Sclerotinia sclerotiorum* apothecial presence in soybean (*Glycine max*) fields. Plant Dis. 102:73–84.

Willbur, J. F., Fall, M. L., Byrne, A. M., Chapman, S. A., McCaghey, M. M., Mueller, B. D., et al. 2018b. Validating *Sclerotinia sclerotiorum* apothecial models to predict sclerotinia stem rot in Soybean (*Glycine max*) fields. Plant Dis. 102:2592–2601.

Willbur, J., McCaghey, M., Kabbage, M., and Smith, D. L. 2018c. An overview of the *Sclerotinia sclerotiorum* pathosystem in soybean: impact, fungal biology, and current management strategies. Trop. Plant Pathol. Available at: http://link.springer.com/10.1007/s40858-018-0250-0 [Accessed December 10, 2018].

Wood, S. N. 2017. *Generalized additive models: an introduction with R*. 2nd ed. Chapman and Hall/CRC. Available at: https://www.taylorfrancis.com/books/9781498728348 [Accessed June 7, 2022].

Xu, X., Madden, L. V., Edwards, S. G., Doohan, F. M., Moretti, A., Hornok, L., et al. 2013. Developing logistic models to relate the accumulation of DON associated with Fusarium head blight to climatic conditions in Europe. Eur. J. Plant Pathol. 137:689–706.

SI APPENDIX

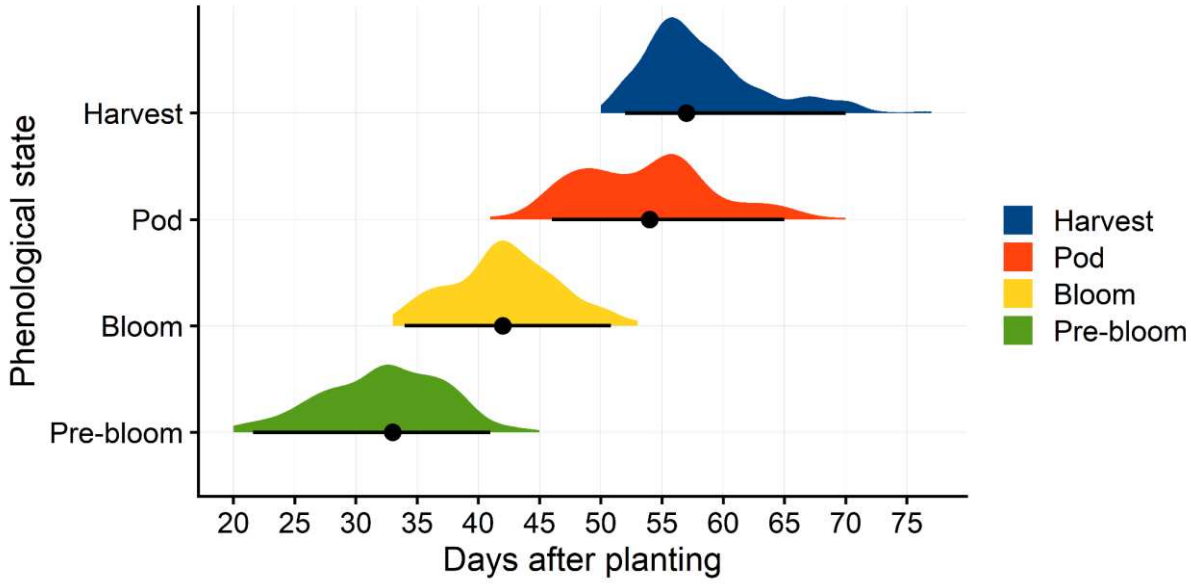

Fig. S1. Kernel density estimates of days after planting for each observed phenological stage in surveyed snap bean fields in New York, USA. Pre-bloom = plants in the vegetative stages; Bloom = at least 50% of plants have at least one open blossom; Pod = developing (immature) pods; Harvest = most pods are mature, and the field is harvestable (Shah et al. 2019a). Points represent the median of the distribution while error bars represent the 95% percentile intervals.

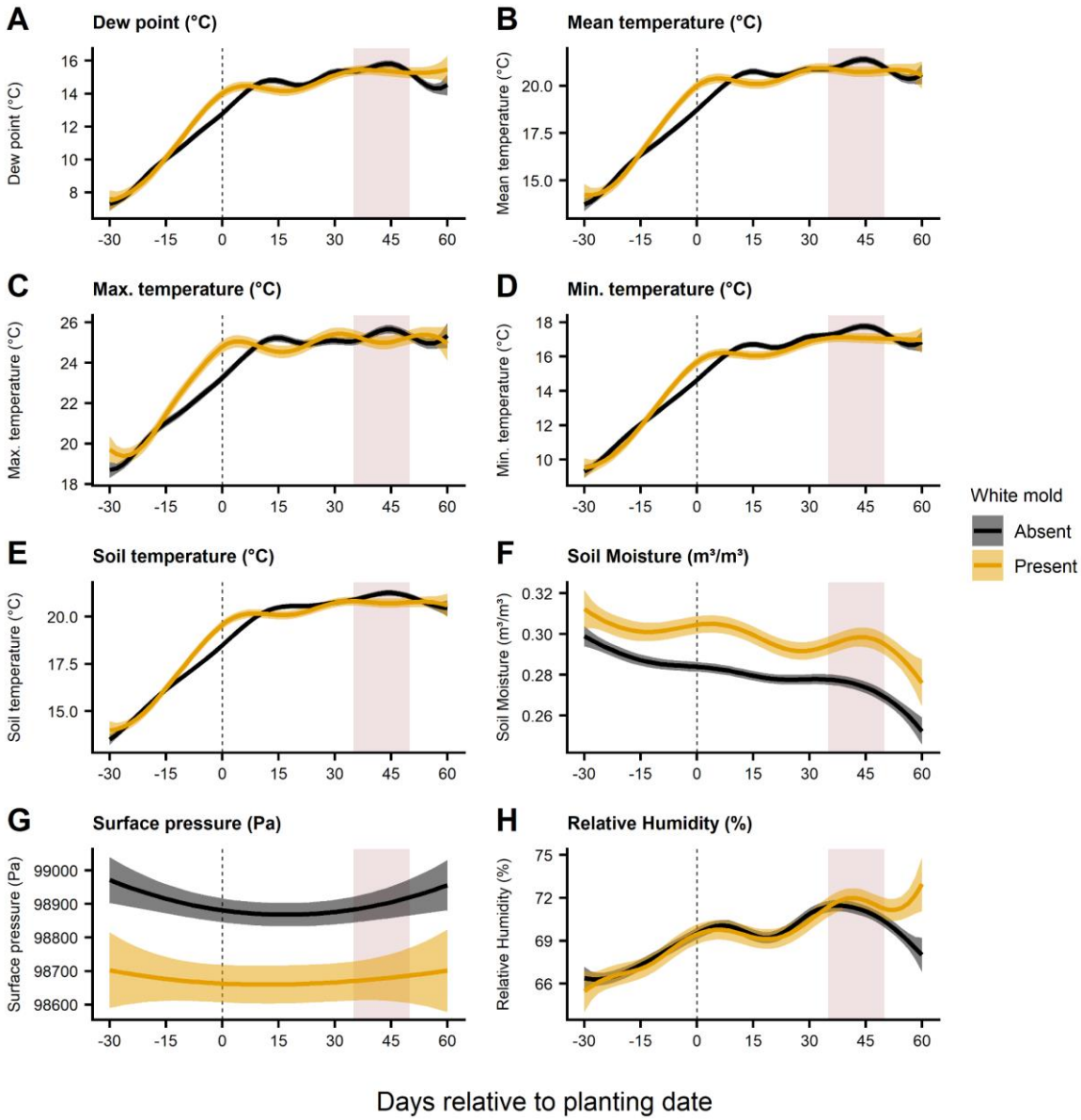

Fig. S2. Functional means (solid lines) and 95% confidence intervals (shaded ribbon) for weather variables associated with snap bean fields in New York, USA in which white mold was present or absent. Functional curves were estimated for: (**A**) Daily mean dew point temperature at 2 meters above ground (2 m); (**B**) Daily mean temperature at 2 m; (**C**) Daily maximum temperature at 2 m; (**D**) Daily minimum temperature at 2 m; (**E**) Daily mean soil temperature within the upper 0 to 7 cm; (**F**) Daily mean soil moisture within the upper 0 to 7 cm; (**G**) Daily mean surface pressure; (**H**) Daily mean relative humidity at 2 m. The rectangular shaded area between 35 and 50 days represents the snap bean bloom (flowering) period.

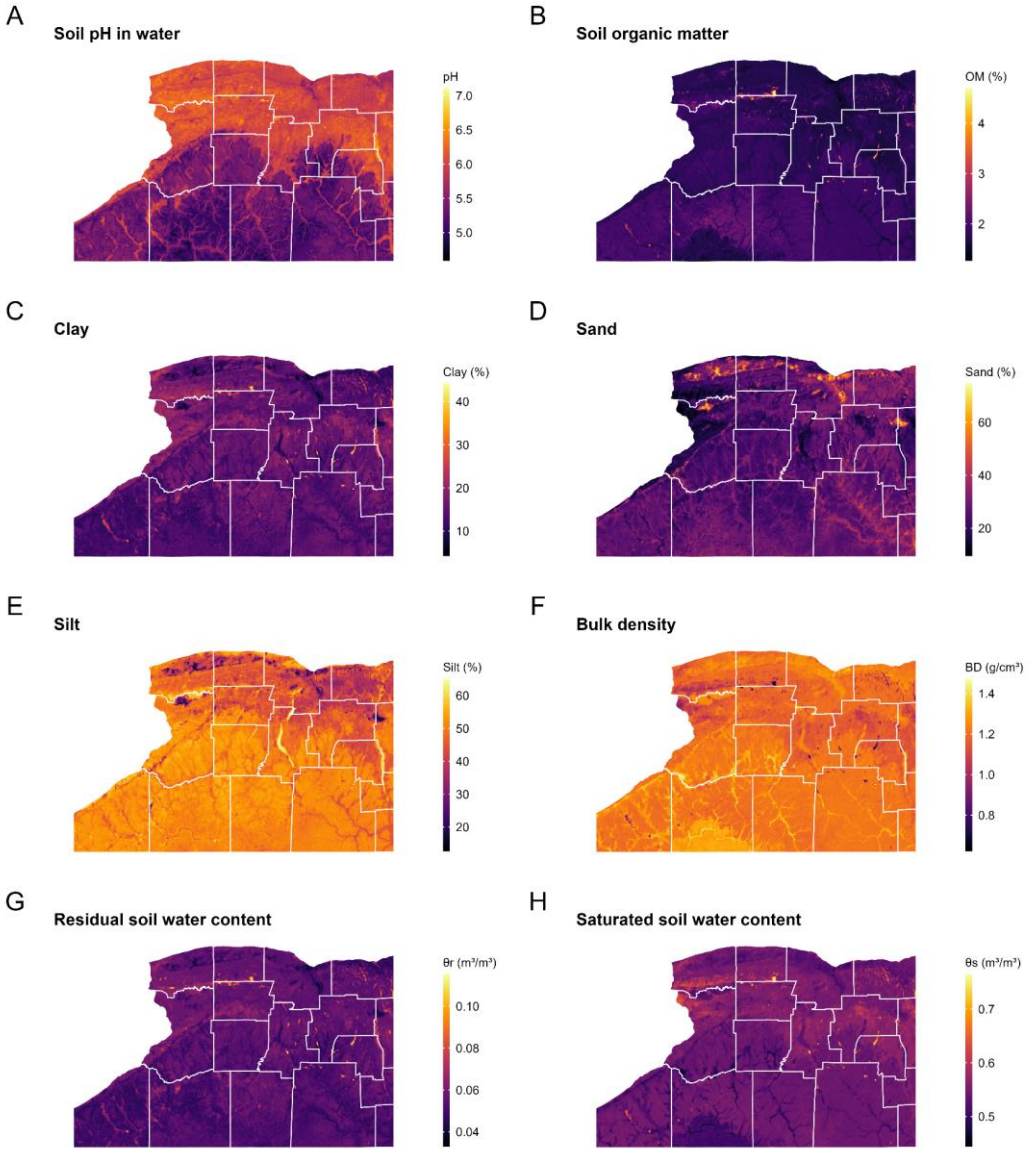

Fig. S3. Gridded spatial distribution of soil properties data retrieved from the POLARIS database over central and western New York, USA.

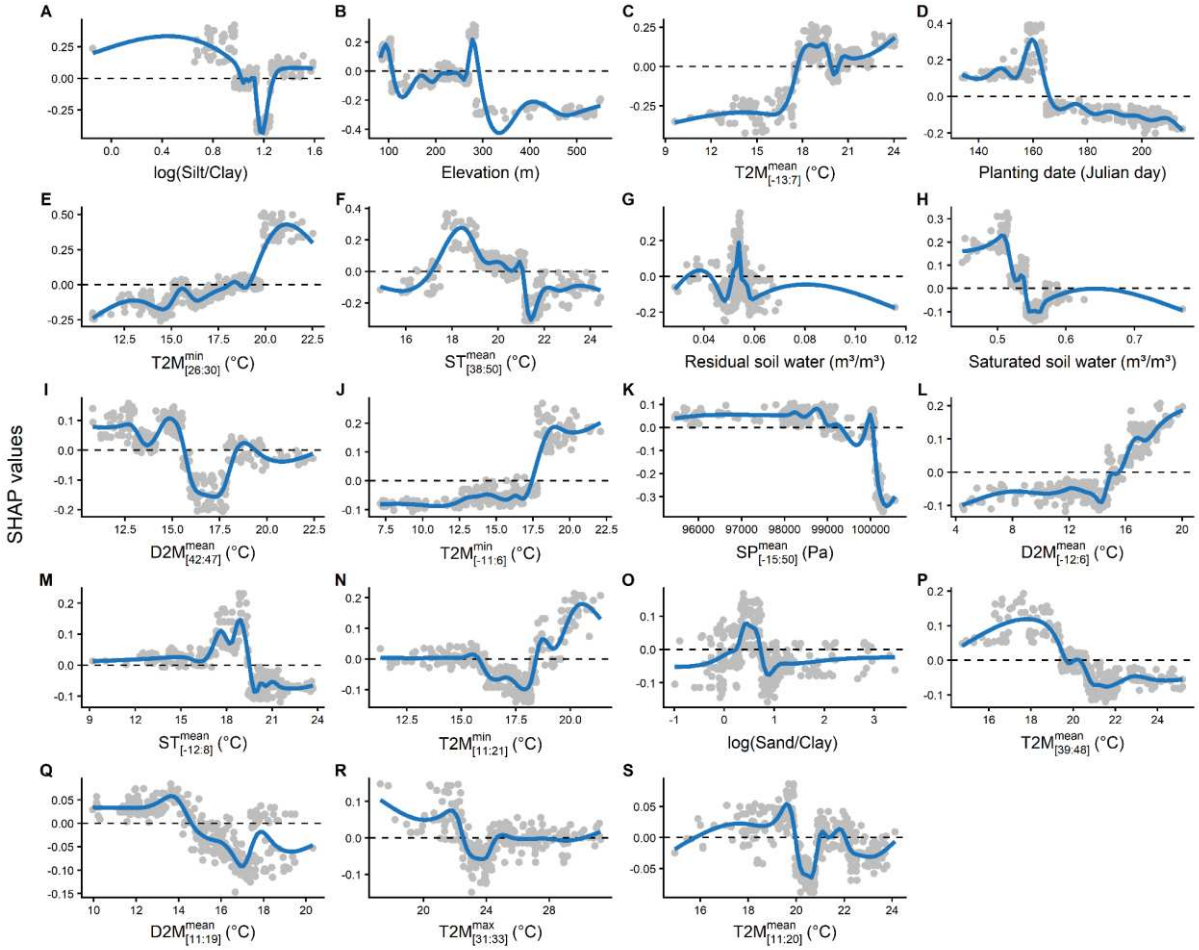

Fig. S4. SHapley Additive exPlanations (SHAP) dependence plots for the 11th to 29th most important environmental variables in an XGBoost model fitted to the presence of white mold in snap beans. Higher SHAP values are associated with an increased probability of white mold being present in the field. Solid lines represent cubic spline smooth curves fitted to the SHAP values. Points represent SHAP values estimated at each observed predictor value in the data (*x* axis). SM = soil moisture within the upper 0 to 7 cm; T2M = air temperature 2 m above ground; ST = soil temperature within the upper 0 to 7 cm; RH = air relative humidity 2 m above ground; SP = surface pressure; D2M = dew point temperature 2 m above ground. For the weather-based variables, the superscript indicates the type of daily value (mean, minimum, or maximum) that was averaged over the period (in days) relative to planting indicated in the subscript (*e.g.*, [40:47] is the period from 40 to 47 days after planting).

Table S1: List of environmental (weather and soil) variables with their respective daily summaries and abbreviated symbols.

Type	Variable[a]	Daily Summary[b]	Window[c]	Abbreviation/ symbol[d]	Unit[e]
Weather	Dew point temperature (2 m AS)	Average	-12 to 6	$D2M^{mean}_{[-12:6]}$	°C
Weather	Dew point temperature (2 m AS)	Average	11 to 19	$D2M^{mean}_{[11:19]}$	°C
Weather	Dew point temperature (2 m AS)	Average	42 to 47	$D2M^{mean}_{[42:47]}$	°C
Weather	Air temperature (2 m AS)	Average	-13 to 7	$T2M^{mean}_{[-13:7]}$	°C
Weather	Air temperature (2 m AS)	Average	11 to 20	$T2M^{mean}_{[11:20]}$	°C
Weather	Air temperature (2 m AS)	Average	39 to 48	$T2M^{mean}_{[11:20]}$	°C
Weather	Air temperature (2 m AS)	Maximum	-15 to 7	$T2M^{max}_{[-15:7]}$	°C
Weather	Air temperature (2 m AS)	Maximum	12 to 19	$T2M^{max}_{[12:19]}$	°C
Weather	Air temperature (2 m AS)	Maximum	31 to 33	$T2M^{max}_{[31:33]}$	°C
Weather	Air temperature (2 m AS)	Maximum	40 to 47	$T2M^{max}_{[40:47]}$	°C
Weather	Air temperature (2 m AS)	Minimum	-11 to 6	$T2M^{min}_{[-11:6]}$	°C
Weather	Air temperature (2 m AS)	Minimum	11 to 21	$T2M^{min}_{[11:21]}$	°C
Weather	Air temperature (2 m AS)	Minimum	26 to 30	$T2M^{min}_{[20:30]}$	°C
Weather	Air temperature (2 m AS)	Minimum	39 to 50	$T2M^{min}_{[39:50]}$	°C
Weather	Surface pressure	Average	-15 to 50	$SP^{mean}_{[-15:50]}$	Pa
Weather	Soil temperature (0 to 7 cm)	Average	-12 to 8	$ST^{mean}_{[-12:8]}$	°C
Weather	Soil temperature (0 to 7 cm)	Average	13 to 23	$ST^{mean}_{[13:23]}$	°C
Weather	Soil temperature (0 to 7 cm)	Average	38 to 50	$ST^{mean}_{[38:50]}$	°C
Weather	Soil moisture (0 to 7 cm)	Average	-15 to 50	$SM^{mean}_{[-15:50]}$	m^3/m^3
Weather	Relative humidity (2 m AS)	Average	40 to 50	$RH^{mean}_{[40:50]}$	%
Soil	Soil pH in water (0 to 5 cm)	-	-	-	-
Soil	Soil organic matter (0 to 5 cm)	-	-	-	%
Soil	Clay (0 to 5 cm)	-	-	-	%
Soil	Sand (0 to 5 cm)	-	-	-	%
Soil	Silt (0 to 5 cm)	-	-	-	%
Soil	Bulk density (0 to 5 cm)	-	-	-	g/cm^3
Soil	Residual soil water content (0 to 5 cm)	-	-	-	m^3/m^3
Soil	Saturated soil water content (0 to 5 cm)	-	-	-	m^3/m^3
-	Elevation	-	-	-	m
-	Planting date	-	-	-	Julian day

[a] Daily average relative humidity was derived from daily average temperature at 2 m above ground (in Kelvin) and daily average dewpoint temperature at 2 m above ground (see Material and Methods); AS = above surface; Log-transformations of the ratios between sand and clay [log(sand/clay)] and silt and clay [log(silt/clay)] were used as predictors in the XGBoost model instead of the original sand, silt, and clay values.

[b] Soil-related variables were time-invariant, hence no summarization was needed.

[c] Days relative to planting.

[d] For the weather-based variables, the superscript indicates the type of daily value (mean,

minimum, or maximum) that was averaged over the period (in days) relative to planting indicated in the subscript (e.g., [40:47] is the period from 40 to 47 days after planting).
[e] The native unit of T2M and D2M from the ERA5-Land database was Kelvin. Values were converted to degrees Celsius (°C) by subtracting 273.15.

Table S2: Hyperparameter values used in the final XGBoost model. The descriptions provided were adapted from the documentation of the functions in the dials (Kuhn and Frick 2022) and in the xgboost (Chen et al. 2022) packages.

Hyperparameter[a]			
dials	**xgboost**	**Description[b]**	**Values[c]**
mtry()	colsample_bynode	Number of predictors that are randomly sampled at each split when creating tree models.	12
trees()	nrounds	Total number of trees in the XGBoost model (default: 15).	2000
min_n()	min_child_weight	Minimum number of data points in a node required for the node to be split further (default: 1).	4
tree_depth()	max_depth	Maximum depth of the tree. (default: 6).	3
learn_rate()	eta	The rate at which the boosting algorithm adapts from iteration-to-iteration. This hyperparameter weighs how much of the error is considered when adding new trees during the learning process (default: 0.3).	0.055
loss_reduction()	gamma	The minimum reduction in the loss function to make a further split on a leaf node of the tree (default: 0).	1.412

[a] Argument names as used in the dials package (used in the tidymodels framework as implemented in this study) and the corresponding argument name in the xgboost package.
[b] Default values as specified for the boost_tree() function of the parsnip package (Kuhn and Vaughan 2022), which is part of tidymodels framework.
[c] Tuned hyperparameter values.

SI References

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. 2022. *xgboost: Extreme Gradient Boosting*. Available at: https://CRAN.R-project.org/package=xgboost.

Kuhn, M., and Frick, H. 2022. *dials: Tools for Creating Tuning Parameter Values*. Available at: https://CRAN.R-project.org/package=dials.

Kuhn, M., and Vaughan, D. 2022. *parsnip: A Common API to Modeling and Analysis Functions*. Available at: https://CRAN.R-project.org/package=parsnip.

Shah, D. A., Dillard, H. R., and Pethybridge, S. J. 2019. Identification of factors associated with white mould in snap bean using tree-based methods. Plant Pathology. 68:1694–1705.

GENERAL CONCLUSIONS

CHAPTER 1

Based on two Bayesian models we were able to find associations between climatic variables and citrus Huanglongbing prevalence in Minas Gerais. The most critical variables were mean temperature during the dry season and mean annual wind speed. The temperature during the dry season was inversely associated with HLB risk, while wind speed was positively associated.

CHAPTER 2

The time to the onset of soybean rust outbreaks in soybean commercial fields in southern Brazil was associated with normal Oceanic Niño Index (ONI; anomalies in normal temperature within the Niño 3.4 region in the Pacific Ocean). A positive association was found, in which higher values of ONI should increase the risk of early outbreaks. Moreover, higher values of ONI (ONI > 0.5) indicate El Niño events, therefore, El Niño events are expected to trigger the early onset of soybean rust epidemics in southern Brazil.

CHAPTER 3

A complete framework comprising weather and soil property data obtention, downscaling and processing data, extraction of candidate variables from weather time series, fusion of datasets, and machine learning interpretation was developed. This framework was applied and validated using a case study on studying the environmental effect on white mold prevalence in snap bean fields in New York, United States. Results indicated that seasonal soil moisture, temperature from different periods of the growing season (including around planting dates), soil organic matter, post-planting soil temperature, relative humidity during flowering, soil density, and soil pH are the most influential environmental variables for white mold in snap beans.