ABSTRACT

OJWANG', AWINO MAUREIQ EDITH. Network Models for the Dispersal of *Pseudoperonospora cubensis* and Spread of Cucurbit Downy Mildew in the Eastern United States. (Under the direction of Dr. Alun Lloyd and Dr. Peter Ojiambo).

Pseudoperonospora cubensis is a long-distance aerially dispersed plant pathogen that causes cucurbit downy mildew. Cucurbit downy mildew (CDM) is a foliar disease that affects the cucurbits. CDM has an annual spread pattern from southern Florida and the Gulf of Mexico to the northeast United States. It is unclear whether this is a spatial spreading process or a climate-driven process. Based on the assumption that CDM spread is influenced by the connectivity of cucurbit fields in time and space, static and dynamic networks were constructed to provide quantitative information on the actual spatiotemporal CDM dynamics. The static networks were sensitive to the choice of parameters and thresholds for construction and identified the potential *P. cubensis* transmission pathways. The dynamic networks facilitated visualization of the time-dependent evolving prominent pathways of pathogen dispersal and identified areas that may act as sources and promote CDM spread. The results provide a framework for understanding the role of network connectivity in predicting CDM spread at the landscape level. When complemented with disease scouting, these results provide valuable information on the effective use of resources when monitoring new CDM outbreaks in the eastern United States.

Cucurbit downy mildew is currently managed using fungicides which must be applied promptly. A platform - linking monitoring, prediction, and communication of the risk of CDM outbreak - has helped growers initiate fungicide treatments in the eastern United States. However, this platform is expensive to maintain, and the resources are often limited. Therefore, knowing where and when to monitor could reduce the costs associated with scouting for disease. In addition, knowing where to treat could slow down the invasion process. A combination of centrality

measures, frequency of infection, and probability of field infection identified highly connected locations for disease surveillance and management in Maryland, North Carolina, Ohio, South Carolina, and Virginia. Removing nodes (representing these locations) limited the risk of disease spread based on a dynamic network model incorporating a power-law function for pathogen dispersal. These locations may inform strategies for controlling CDM in the eastern United States.

The power-law and logistic phenomenological models have been used in plant pathology to describe the spatial and temporal rate of change of disease intensity, respectively. Two assumptions are that pathogen dispersal is isotropic, and there is one inoculum source. These models were extended to incorporate anisotropy and multiple inoculum sources using an estimation framework suited to describing disease spread in space and time. Based on the analysis of epidemic data from 2008 to 2010, there was a small but consistent reduction in errors associated with incorporating anisotropy into the model regardless of the number of sources and a reduction in errors in specific years associated with incorporating an alternate inoculum source in Texas. However, there was no reduction in errors associated with incorporating an alternate inoculum source in Ohio. These results strongly suggest that the initial inoculum source for CDM outbreaks in the continental United States is primarily from overwintering sources in the southern United States.

Network Models for the Dispersal of *Pseudoperonospora cubensis* and Spread of Cucurbit Downy Mildew in the Eastern United States

by
Awino Maureiq Edith Ojwang'

A dissertation submitted to the Graduate Faculty of Science
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Biomathematics

Raleigh, North Carolina
2021

APPROVED BY:

Dr. Alun L. Lloyd Committee Co-Chair	Dr. Peter Ojiambo Committee Co-Chair
Dr. Ross Meentemeyer	Dr. Charles Smith

DEDICATION

To Dr. Ayoma Ojwang’

BIOGRAPHY

Awino Maureiq Edith Ojwang' was born in Kajiado, Kenya on March 23, 1987. After graduating from Asumbi Girls High School in Homabay, Kenya, she attended Jomo Kenyatta University of Agriculture and Technology, graduating with a Bachelor of Science in mathematics and computer science. In 2010, Maureiq decided to return to school and received a Master of Science in Bioinformatics from the University of Nairobi. After graduation, Maureiq first took a teaching assistant position to Dr. Sarah Schaack, where she traveled across east Africa teaching bioinformatics and genomics workshops in Universities and research institutions. After this, Maureiq took a part-time lecturer position at the Centre for Biotechnology and Bioinformatics, University of Nairobi, and a Bioinformatician position at the Biosciences in eastern and central Africa (BecA)-ILRI Hub in Kenya. Looking for ways to combine her interest in mathematics and biology, she decided to attend North Carolina State University to work on cucurbit downy mildew under Dr. Alun Lloyd and Dr. Peter Ojiambo. Maureiq hopes to continue her work to help develop more efficient control strategies for epidemics.

ACKNOWLEDGMENTS

I am deeply thankful to my advisors for their guidance and support over the past five years. I want to thank Dr. Peter Ojiambo for his research support, writing, and encouragement to continue even when I was overwhelmed with research. I am very grateful to Dr. Alun Lloyd for his research support and for ensuring that I had the financial support to finish my fifth year. A special thank you to Dr. Charles Smith (Statistics department) and Dr. Ross Meentemeyer (Geospatial Analytics), for sharing your expertise with me and for your guidance during the committee meetings. I would also like to thank Dr. Kevin Gross, who continually challenged me to think about science, its implications, and its impact. I am very thankful to Dr. Trevor Ruiz, Dr. David Gent, Dr. Sharmodeep Bhattacharyya, and Dr. Shirshendu Chatterjee at Oregon State University, who contributed significantly to my work.

I am also thankful for the support system I had while at NCSU, both inside and outside my program. Thanks to Dr. Josephine Birungi, Dr. Sarah Schaack, Dr. Kevin Flores, and Dr. Jacqueline Hughes-Oliver for your invaluable support and help over the years. I am also grateful to Dr. Sheila Okoth, Dr. Patrick Weke, and the late Dr. James Ochanda for pushing me towards biomathematics. Thank you to all of my friends that have helped me keep some balance in my life. In particular, many thanks to Dr. Doris Sande, Pat Cate, and the late Mary-Ann Cate, who provided for all my needs when I came to the United States. I also wish to extend my thanks to Cindy Bell and Chuck Mays for providing a suitable living environment. Thank you to Dr. Brandon Hollingworth for helping me a lot, especially during this final push. Also, thank you to Shelagh and Pete, Katie, Anna, and Marian for helping me during my first year. Finally, thank you to Yeng, Amanda, Mitchel, John, Marco, Julian, Praachi, Annabel, Evan, and everyone else for being such good friends over the years.

Most importantly, thank you to my God and my family. I would not have made it without you. You have supported me through everything. To my father, the late Dr. Ayoma Ojwang', I will never be able to express how grateful I am for the foundation you laid for me. To my mom, Mrs. Rose Ojwang', thank you so much for all your prayers and support through all this. Thank you to my siblings, Mercy Ojwang', the late Susan Ojwang', William Ojwang', and Lameck Ojwang' for your prayers and unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

CHAPTER 1

Introduction

1.1 Cucurbits and cucurbit downy mildew

Cucurbit downy mildew (CDM) is a foliar disease that affects the cucurbits. Cucurbits are a large group of gourd-like plants made up of 118 genera with 825 species (Lebeda and Cohen, 2011). These include cucumber (*Cucumis sativus* L.), melon (*Cucumis melo*), squash (*Cucurbita* spp.), giant pumpkin (*Cucurbita maxima*), watermelon (*Citrullus lanatus*), among others. Cucurbits are economically important crops, having edible fruits, leaves, flowers, and seeds, which are sources of cooking oil (Zitter et al., 1998). Cucurbits are cultivated globally but mainly in the tropics (Cutler and Whitaker, 1961). The main cucurbits grown commercially in the United States include melons, squash, pumpkins, and cucumbers. Nationally, melons, squash, and pumpkins are grown on over 27,046, 140,380, and 187,125 hectares, respectively (NASS, 2017). However, cucumbers are grown on more than 239,311 hectares (NASS, 2017), resulting in an average production of 1 billion kilograms making the United States the third-largest producer of cucumbers worldwide (Neufeld et al., 2018). The states located in the southeast (e.g., Florida, Georgia, Alabama, South Carolina, North Carolina, Mississippi, and Tennessee) and the northeast United States have the largest commercial cucurbits production (Figure 1.1).

Cucurbits are sensitive to frost and thus only grow below the 30° latitude during the winter months in the continental United States. These areas are subject to warm and humid weather conducive for the development of cucurbit downy mildew. Cucurbit downy mildew occurs annually in the eastern United States, where it causes devastating losses of cucurbits in the absence of adequate disease management (Ojiambo et al., 2011). The disease causes different symptoms in different hosts (Figure 1.2) and is caused by the oomycete *Pseudoperonospora cubensis.* The pathogen infects 60 species and 20 genera of cucurbits, with the *Cucumis*, *Cucurbita*, and *Citrullus* genera being important hosts. The *Cucumis* susceptible hosts include eight wild species and two

cultivated species: cucumber *(C. sativus* L.) and muskmelon (*C. melo* L.) (Lebeda and Cohen, 2011).

1.2 *Pseudoperonospora cubensis*

Pseudoperonospora cubensis is an obligate biotrophic pathogen that belongs to kingdom Chromista, subdivision Peromosporomycotina, class Oomycetes, order Peronosporales, and family Peronosporaceae (Lebeda and Cohen, 2011). *P. cubensis* was known as *Peronospora cubensis*, *Plasmopara cubensis*, and *Peronoplasmopara cubensis* in the past (Dick, 2001). The pathogen reproduces sexually through the production of oospores or asexually through the production of sporangia (Lebeda and Cohen, 2011). The sporangia are the main infective inoculum (Thomas, 1996). They are light grey to deep purple (Thomas, 1996), 20 to 40 × 14 to 25 μm in diameter, and are oval-shaped, having a papilla at the distal end (Lebeda and Cohen, 2011). The sporangia develop at the tips of hyaline sporangiophores (Choi et al., 2005).

Sporangiophores arise from stomatal openings in sets of two to six (Choi et al., 2005) on the abaxial surface (underside) of leaves (making it easy for the sporangia to be dislodged from the tips of sporangiophores and dispersed by wind). The sporangiophores twist - when there is a decrease in relative humidity - and release sporangia in the air (Lange et al., 1989). The aerially borne sporangia are dispersed via wind and travel long distances up to 1,000 km from the initial inoculum source (Ojiambo and Holmes, 2011). The survival of sporangia in the air is affected by temperature, humidity, and solar radiation (Kanetis et al., 2010; Thomas, 1996), with the latter having the most effect on sporangia survival (Kanetis et al., 2010). Depending on weather conditions, the viability of the aerially borne sporangia ranges from one to sixteen days (Cohen, 1981), with cloudy days prolonging sporangia survival.

Dissemination on susceptible hosts occurs during rainfall when sporangia in the air are deposited on the hosts. Infectivity and subsequent germination of sporangia require conducive environmental conditions, i.e., optimum temperature, humidity, leaf wetness, and absence of light. Infection requires optimum temperature between 15°C and 20°C (Cohen and Rotem, 1969), with infection at 30°C possible (Arauz et al., 2010). Infection also requires a minimum duration of two hours of leaf wetness (Cohen, 1977). After infection of susceptible hosts, within five to seven days, sporangiophores develop inside the host and emerge through stomatal openings (Choi et al., 2005) when the relative humidity > 90% (Lebeda and Cohen, 2011). *P. cubensis* causes lesions on the adaxial surface of leaves which vary in size, shape, and color in different hosts. For example, the lesions are angular in cucumber because leaf veins restrict them. However, in watermelons, cantaloupes, and squash, the lesions are circular and irregular because this restriction is absent (Figure 1.2). After the initial infection, lesions expand, mature, sporulate, and may become necrotic. The lesions continue to expand after some days, coalesce, kill leaves, eventually killing the entire plant (Lebeda and Cohen, 2011). Conducive environmental conditions influence the timing and duration of sporulation. For example, low temperatures delay the onset of sporulation but increase the duration that a lesion will sporulate (Cohen, 1977). In addition, sporulation can occur in as little as four to five days on mature lesions under conducive conditions (Lebeda and Cohen, 2011). The absence of light also affects sporulation, with heavy sporulation occurring after long periods of darkness (Cohen and Eyal, 1977), i.e., at least six hours (Cohen, 1977).

The incubation period of *P. cubensis* ranges from four to twelve days (Cohen, 1977), sporangiophores develop in five to seven days, and a new infection cycle starts seven to ten days under field conditions (Lebeda and Cohen, 2011). *P. cubensis* results in epidemics where inoculum is produced by plants previously infected during the same epidemic that season. However, the

initial inoculum source is a subject of intense research (Neufeld et al., 2018). In the continental United States, *P. cubensis* is generally assumed to overwinter in cultivated or wild cucurbits in frost-free areas < 30° latitude, i.e., in southern Florida and along the Gulf of Mexico (Bains and Jhooty, 1976). Thus, the onset of disease in areas > 30° latitude requires aerial dispersal from these overwintering sources (Nusbaum, 1944). The protected greenhouses could also be potential inoculum sources (Neufeld et al., 2018; Ojiambo and Holmes, 2011).

1.3 Disease management

Successful disease management requires that control measures be implemented just before the first detection of cucurbit downy mildew in cucurbit fields. The level and extent of disease management vary depending on the planting type, i.e., commercial, research, disease monitoring (sentinel plots), or private (home gardens) (Zitter et al., 1998). Disease control using fungicides is usually limited for home gardens but needs to be implemented in commercial fields to avoid complete crop loss (Neufeld et al., 2018). From the 1950s to 2004, cucurbit downy mildew was mainly managed using host resistance. In the United States, the disease was managed in commercial cultivars by the recessive *dm*1 gene (Holmes and Thomas, 2006; Holmes et al., 2006). A change in the pathogen population structure occurred in 2004, resulting in a breakdown of host resistance and subsequently widespread losses of cucumber crops in the United States (Holmes and Thomas, 2006; Holmes et al., 2015). Since then, growers have relied on the use of fungicides to manage the disease. This management requires fungicide sprays every five to seven days for cucumbers and seven to ten days for other cucurbits (Hausbeck and Cortright, 2009), and early onset of the disease can require up to eleven sprays to control the disease (Ojiambo et al., 2010). Effective management using fungicides requires the precise timing of the initial spray. Given the

rapid development of CDM, prediction and timing of the initial spray can significantly reduce the total number of subsequent sprays, fungicide costs, and resistance to fungicides. In the United States, a CDM forecasting system is available to help growers decide when to apply the first spray (Holmes et al., 2004; Ojiambo et al., 2011).

The United States Department of Agriculture Pest Information Platform for Extension and Education (USDA PIPE) developed the cucurbit downy mildew forecasting system, CDM ipmPIPE, in 2008. The forecasting system uses aerobiological models to forecast CDM and provides growers with a real-time epidemic status of CDM spread in the United States (Ojiambo et al., 2011). The forecasts are based on

1. Disease outbreaks in sentinel and non-sentinel plots reported to http://cdm.ipmpipe.org.
2. The projected sporangia transport routes based upon prevailing wind conditions.
3. The right weather conditions for sporangia deposition, infection, and cucurbit downy mildew development within the neighboring areas.

Growers receive customized reports (forecasts) and start fungicide treatments on their cucurbit fields when there is a threat or risk of infection (Ojiambo et al., 2011). Information from the CDM ipmPIPE has been fundamental in understanding the CDM spread in the United States. For example, Ojiambo and Kang (2013) examined patterns of time to disease outbreak in the United States using Bayesian hierarchical spatially structured frailty models and showed that the risk of disease outbreak was high in the mid-Atlantic region (Ojiambo and Kang, 2013). Similarly, based on records of disease outbreaks, Ojiambo et al. (2017) showed that the position of the epidemic wavefront became exponentially more distant with time, and epidemic velocity increased linearly with distance. Further, the authors observed that the final epidemic extent was correlated to the size of the initial epidemic area, and efforts to reduce the initial epidemic area can be useful

in mitigating focus expansion and subsequent spread of the disease in more northern states (Ojiambo et al., 2017).

1.4 Rationale and justification

The pattern of cucurbit downy mildew spread from southern Florida and the Gulf of Mexico to the northern states is consistent. We understand how the epidemic moves based on the observation data. However, we do not understand some of the dynamics of how this happens in space and time and how this information can help us better manage the disease. If we assume that the field connectivity influences CDM spread in time and space, we can formulate network models to describe the CDM spread similar to other pathosystems (Sanatkar et al., 2015; Sutrave et al., 2012; Gent et al., 2019). Networks can help us understand the signatures of disease spread, visualize potential pathways of spread, and provide a framework for understanding the role of network connectivity in predicting disease spread at the landscape level. Furthermore, the successful use of networks to describe soybean rust caused by *Phakopsora pachyrhizi* and hop downy mildew caused by *Podosphaera macularis* motivated the use of network models to describe the dispersal of *Pseudoperonospora cubensis* and the spread of CDM in the eastern United States.

The CDM forecasting system has been useful in guiding growers on initiating fungicide treatments in the eastern United States (Ojiambo et al., 2011). However, this system is expensive to maintain, and the resources are often limited for monitoring and reporting disease outbreaks. Therefore, knowing where and when to monitor for the disease could reduce economic costs if the goal is to collect data efficiently with the least possible resources. Further, knowing where to treat could slow down the invasion process during the growing season. Disease outbreaks at the county scale may also be influenced by selective removal of some severely infected fields. Node centrality

measures can characterize the role of a node in a network based on its connection topology. For example, a study of soybean rust in the United States identified key geographical nodes for sampling to forecast disease spread (Sanatkar et al., 2015). Further, combining these measures with the frequency of infection and probability of field infection can be used to identify the important locations for monitoring and management. In this dissertation, various strategies to reduce the number of monitoring sites are explored and discussed to reduce the cost involved while still maintaining disease prediction accuracy.

Disease spread dynamics in a spatial context are influenced by anisotropy and multiple inoculum sources, among other factors. Existing phenomenological models, e.g., the model used by Ojiambo et al. (2017), while still informative, do not account for anisotropy and multiple foci. A few phenomenological models account for anisotropy for relatively short dispersal of seeds, pollen, and pathogen propagules and single pathogen generation or dispersal event, but do not consider anisotropy in epidemic spread in time (Rieux et al., 2014; Soubeyrand et al., 2007; van Putten et al., 2012). Thus, there is a need to expand these models to incorporate anisotropy and establish if this characteristic of pathogen dispersal results in better estimates of disease spread in time and space. In addition, the presence of alternative sources of *P. cubensis* inoculum driving epidemic spread has been proposed (Ojiambo and Holmes, 2011). However, this hypothesis has not been tested using empirical data. In general, dispersal is usually anisotropic for long-distance dispersed pathogens such as *P. cubensis*, and anisotropy may be due to landscape features, host availability, and weather conditions (Taylor et al., 1993). Thus, there is a need to account for multiple sources of initial inoculum since the location and strength of different sources of initial inoculum can impact the extent to which disease-free fields will get infected. Based on the above considerations, the overall goal of this dissertation is to characterize the network structures of

CDM spread in the eastern United States and develop a generalized framework that accounts for multiple inoculum sources and anisotropy for disease spread in time and space. The specific objectives are to:

1. Develop static and dynamic networks to describe the dispersal of *P. cubensis* and the spread of cucurbit downy mildew using historical data (Chapter 2).
2. Develop networks to identify the most important and highly connected locations critical in the spread of cucurbit downy mildew (Chapter 3).
3. Extend the work of Ojiambo et al. (2017) and Rieux et al. (2014) with a modified power-logistic model to incorporate anisotropy and multiple inoculum sources (Chapter 4).

References

1. Arauz, L. F., Neufeld, K. N., Lloyd, A. L., and Ojiambo, P. S. 2010. Quantitative models for germination and infection of *Pseudoperonospora cubensis* in response to temperature and duration of leaf wetness. Phytopathology 100:959-967.
2. Bains, S. S., and Jhooty, J. S. 1976. Host-range and possibility of pathological races in *Pseudoperonospora cubensis*-cause of downy mildew of muskmelon. Indian Phytopathol. 29:214-216.
3. Choi, Y. J., Hong, S. B., and Shin, H. D. 2005. A re-consideration of *Pseudoperonospora cubensis* and *P. humuli* based on molecular and morphological data. Mycol. Res. 109:841-848.
4. Cohen, Y., and Rotem, J. 1969. The effects of lesion development, air temperature, and duration of moist period on sporulation of *Pseudoperonospora cubensis* in cucumbers. Israel J. Bot. 18:135-140.
5. Cohen, Y., and Eyal, H. 1977. Growth and differentiation of sporangia and sporangiophores of *Pseudoperonospora cubensis* on cucumber cotyledons under various combinations of light and temperature. Physiol. Plant. Pathol. 10:93-103.
6. Cohen, Y. 1977. The combined effects of temperature, leaf wetness, and inoculum concentration on infection of cucumbers with *Pseudoperonospora cubensis*. Can. J. Bot. 55:1478-1487.
7. Cohen, Y. 1981. Downy mildew of cucurbits. Pages 341-354 in: The downy mildews. D. M. Spencer, ed. Academic Press, Inc., London.
8. Cutler, H., and Whitaker, T. 1961. History and distribution of the cultivated cucurbits in the Americas. Am. Antiq. 26:469-485.

9. Dick, M. W. 2001. Straminipilous Fungi: Systematics of the Peronosporomycetes, Including Accounts of the Marine Straminipilous Protists, the Plasmodiophorids, and Similar Organisms. Springer, Dordrecht, The Netherlands.
10. Ferrari, J. R., Preisser, E. L. and Fitzpatrick, M. C. 2014. Modeling the spread of invasive species using dynamic network models. Biol. Invasions 16:949-960.
11. Gent, D. H., Bhattacharyya, S., and Ruiz, T. 2019. Prediction of spread and regional development of hop powdery mildew: A network analysis. Phytopathology 109:1392-1403.
12. Hausbeck, M. K., and Cortright, B. D. 2009. Evaluation of fungicides for control of downy mildew of pickling cucumber, 2007. Plant. Dis. Manag. Rep. 3: V112.
13. Holmes, G. J., Main, C. E., and Zeever, Z. T. 2004. Cucurbit downy mildew: a unique pathosystem for disease forecasting. Pages 69-80 in: Advances in Downy Mildew Research – Vol. 1. P.T.N. Spencer-Phillips, and M. Jeger, eds. Kluwer Academic Publishers, Dordercht, The Netherlands.
14. Holmes, G. J., and Thomas, C. E. 2006. The history and re-emergence of cucurbit downy mildew (Abstr.). Phytopathology 99: S171.
15. Holmes, G. J., Ojiambo, P. S., Hausbeck, M. K., Quesada-Ocampo, L., and Keinath, A. P. 2015. Resurgence of cucurbit downy mildew in the United States: a watershed event for research and extension. Plant Dis. 99:428-441.
16. Kanetis, L., Holmes, G. J., and Ojiambo, P. S. 2010. Survival of *Pseudoperonospora cubensis* sporangia exposed to solar radiation. Plant Pathol. 59:313-323.
17. Lange, L., Eden, U., and Olson, L. W. 1989. The zoospore of *Pseudoperonospora cubensis*. The causal agent of cucurbit downy mildew. Nordic J. Bot. 8:511-516.

18. Lebeda, A., and Cohen, Y. 2011. Cucurbit downy mildew (*Pseudoperonospora cubensis*)-biology, ecology, epidemiology, host-pathogen interaction and control. Eur. J. Plant Pathol. 129:157-192.

19. Nusbaum, C. J. 1944. The seasonal spread and development of cucurbit downy mildew in the Atlantic coastal states. Plant Dis. 28:82–85.

20. Ojiambo, P. S., and Holmes, G. J. 2011. Spatiotemporal spread of cucurbit downy mildew in the eastern United States. Phytopathology 101:451-461.

21. Ojiambo, P. S., Holmes, G. J., Britton, W., Keever, T., Adams, M. L., Babadoost, M., Bost, S. C., Boyles, R., Brooks, M., Damicone, J., Draper, M. A., Egel, D. S., Everts, K. L., Ferrin, D. M., Gevens, A. J., Gugino, B. K., Hausbeck, M. K., Ingram, D. M., Isakeit, T., Keinath, A. P., Koike, S. T., Langston, D., McGrath, M. T., Miller, S. A., Mulrooney, R., Rideout, S., Roddy, E., Seebold, K.W., Sikora, E. J., Thornton, A., Wick, R. L., Wyenandt, C. A. and Zhang, S. 2011. Cucurbit downy mildew ipmPIPE: a next generation web-based interactive tool for disease management and extension outreach. Online. Plant Health Progress 0411-01-RV.

22. Ojiambo, P. S., Gent, D. H., Quesada-Ocampo, L. M., Hausbeck, M. K., and Holmes, G. J. 2015. Epidemiology and Population Biology of *Pseudoperonospora cubensis*: A Model System for Management of Downy Mildews. Annu. Rev. Phytopathol. 53:223-246.

23. Ojiambo, P. S., Gent, D. H., Mehra, L. K., Christie, D., and Magarey, R. 2017. Focus expansion and stability of the spread parameter estimate of the power law model for dispersal gradients. PeerJ, 6, 1-20.

24. Ojiambo, P. S., Paul, P. A., and Holmes, G. J. 2010. A quantitative review of fungicide efficacy for managing downy mildew in cucurbits. Phytopathology 100:1066-1076.

25. Ojiambo, P. S., and Kang, E. L., 2013. Modeling spatial frailties in survival analysis of cucurbit downy mildew epidemics. Phytopathology 103:216-227.

26. Rieux A., Soubeyrand, S., Bonnot, F., Klein, E. K., Ngando, J. E., Mehl, A., et al. 2014. Long-distance wind-dispersal of spores in a fungal plant pathogen: estimation of anisotropic dispersal kernels from an extensive field experiment. PLoS One. 9:e103225.

27. Sanatkar, M. R., Scoglio, C., Natarajan, B., Isard, S. A., and Garrett, K. A. 2015. History, epidemic evolution, and model burn-in for a network of annual invasion: Soybean rust. Phytopathology 105:947-955.

28. Savory, E. A., Granke, L. L., Quesada-Ocampo, L. M., Varbanova, M., Hausbeck, M. K., Day, B. 2010. The cucurbit downy mildew pathogen *Pseudoperonospora cubensis*. Mol. Plant Pathol. 12:217-226.

29. Soubeyrand, S., Enjalbert, J., Sanchez, A., and Sache, I. 2007. Anisotropy, in density and in distance, of the dispersal of yellow rust of wheat: experiments in large field plots and estimation. Phytopathology 97:1315-1324.

30. Sutrave, S., Scoglio, C., Isard, S., Hutchinson, J. M. S., and Garrett, K. A. 2012. Identifying highly connected counties compensates for resource limitations when evaluating national spread of an invasive pathogen. PLoS One 7:e37793.

31. Thomas, C. E. 1996. Downy mildew. Pages 25-27 in: Compendium of Cucurbit Diseases. T. A. Zitter, D. L. Hopkins, and C. E. Thomas, eds. American Phytopathological Society Press, St. Paul, MN.

32. Thomas, A., Carbone, I., Cohen, Y., Ojiambo, P. S. 2017. Occurrence and distribution of mating types of *Pseudoperonospora cubensis* in the United States. Phytopathology 107:313-321.

33. van Putten, B., Visser, M. D., Muller-Landau, H. C., and Jansen, P. A. 2012. Distorted-distance models for directional dispersal: a general framework with application to a wind-dispersed tree. Methods Ecol. Evol. 3:642-652.
34. Zitter, T. L., Hopkins, D. L., and Thomas, C. E. 1998. Compendium of Cucurbit Diseases. American Phytopathological Society Press, St. Paul, MN.

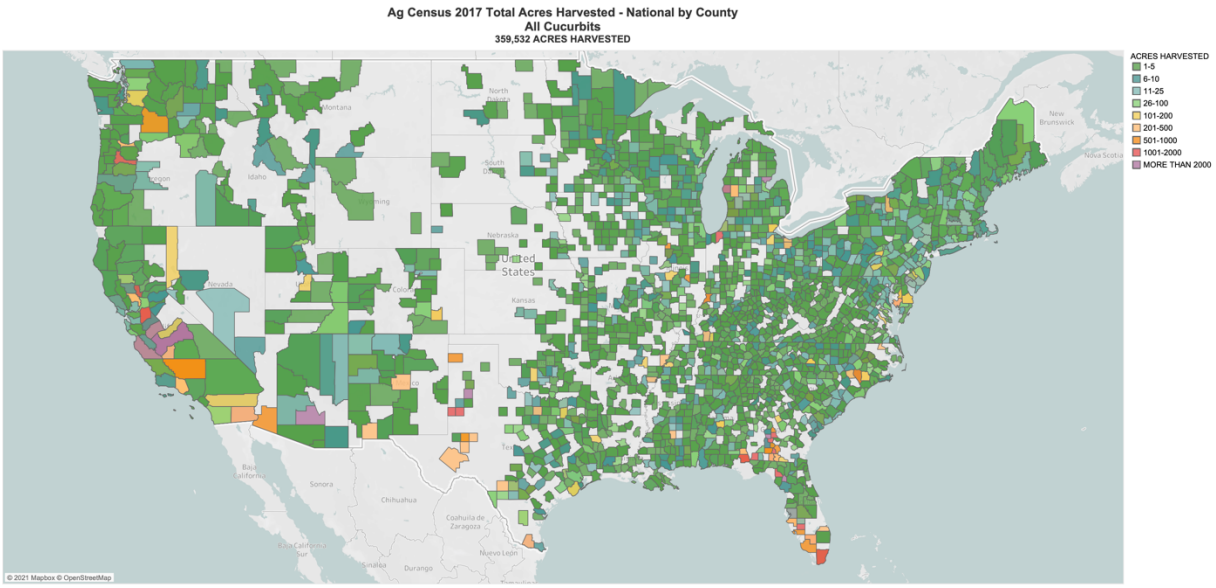


Figure 1.1. The total acres of cucurbits harvested nationally in the United States based on the 2017 USDA agricultural census. Florida, Georgia, Alabama, South Carolina, North Carolina, Mississippi, Tennessee, and California have the largest commercial production of cucurbits.

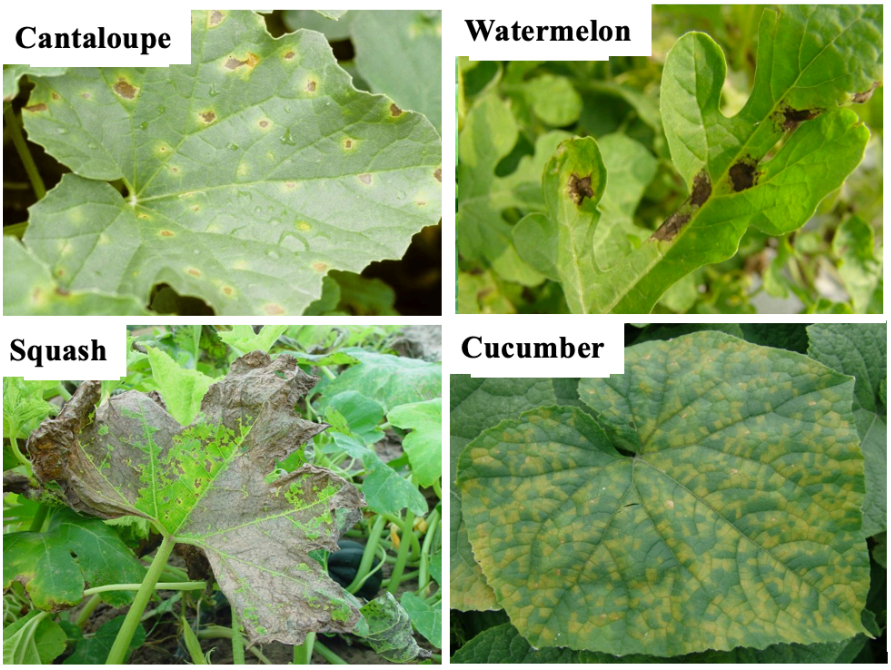


Figure 1.2. Cucurbit downy mildew symptoms. The disease causes different symptoms in different hosts. In cucumber, the lesions are angular in shape because leaf veins restrict them. The cantaloupe, squash, and watermelon lesions are more circular and irregular because this restriction is absent. Photos courtesy of Dr. Gerald J. Holmes.

CHAPTER 2

Modeling Cucurbit Downy Mildew Dynamics in the Eastern United States: A Static and Dynamic Network Analysis

Abstract

Pseudoperonospora cubensis is a long-distance aerially dispersed plant pathogen that causes cucurbit downy mildew. The probability of *P. cubensis* transmission and cucurbit downy mildew spread from infected to disease-free cucurbit growing fields can be expressed in a model that explicitly considers network structure to represent disease spread. Network models were formulated to describe the weekly cucurbit downy mildew spread from overwintering sources in southern Florida to cucurbit fields in northern latitudes in the eastern United States. The fields where cucurbit downy mildew was reported were considered nodes in the networks. Static networks generated using epidemic data recorded from 2008 to 2016 were characterized using properties that define a giant component, density, degree, degree distribution, and centrality. These network properties varied among epidemic years examined. The analysis indicated that nodes in Michigan, Maryland, North Carolina, Ohio, South Carolina, and Virginia were most central in the network and were responsible for spreading cucurbit downy mildew to other fields.

Further, dynamic network models were explored to provide quantitative information on the actual spatiotemporal cucurbit downy mildew dynamics. Early in the cropping season, the transmission probability occurring along links was highest between nodes closest to the initial focus in southern Florida compared to links between nodes elsewhere within the network. As the epidemic progressed in time and space, link probabilities increased for nodes more distant from the node where the initial outbreak was observed. Out-degree values from the dynamic network models were always less than the in-degree, with most of the nodes in the network tending to act as sinks. These results provide a framework for understanding the role of network connectivity in predicting the dispersal of *P. cubensis* at the landscape level and provides useful information on

the effective use of resources to monitor new cucurbit downy mildew outbreaks in the eastern United States.

Keywords: cucurbit downy mildew, static network, dynamic network

1. Introduction

Long-distance dispersal (LDD) occurs when a pathogen's inoculum arrives but does not necessarily establish at a location far away from the original inoculum source (Nathan et al., 2003). Invasive pathogens exhibiting LDD can result in epidemics that are difficult to control (Severns et al., 2019). These pathogens are dispersed via either wind, water, or vectors from initial inoculum sources (point or multiple). They are transmitted over long distances and cause disease away from the source. Such pathogens may generate patterns of disease spread due to long-distance dispersal (Mundt et al., 2009a, 2009b). The annual occurrence and experimental tractability of epidemics caused by LDD pathogens make them excellent model systems for understanding the role of dispersal in disease outbreaks and subsequent spread. Cucurbit downy mildew (CDM) epidemics are an excellent example in this regard. The disease is caused by the oomycete *Pseudoperonospora cubensis*, an aerially dispersed LDD pathogen whose sporangia can be transported over long distances ranging from 700 to 1,000 km. The pathogen exhibits annual recolonization and extinction cycles, generating annual invasions at the continental scale in the U.S. (Ojiambo and Holmes, 2011).

The primary dispersal mechanism for inoculum produced by aerially dispersed plant LDD pathogens is wind. Further, these pathogens capable of long-distance dispersal follow a power-law dispersal gradient resulting in disperse epidemic waves (Ferrandino 1993; Scherm 1996; Ojiambo et al., 2017). The disease gradient of pathogens dispersed some distance away from a source can be described using the phenomenological power-law model (Campbell and Madden, 1990)

$$\frac{dy}{ds} = \frac{-by}{s} \quad (1)$$

or a modified phenomenological power-law (power-logistic model) that allows for calculations at $s = 0$ and also account for limitations of disease

$$\frac{dy}{dr} = \frac{-by(1-y)}{r\ +\ s} \quad (2)$$

where r is an offset parameter (Madden et al., 2007), s is the distance from source s_o, b is the spread parameter, $y = \frac{Y}{M}$ is the infected quantity (where Y represents the disease in absolute units such as the number of lesions, infected leaves or plants, while M represents the total number of plants or plant area that can be infected), and $1 - y$ is the healthy quantity (Madden et al., 2007). Equation 2 is a phenomenological power-logistic model used in a study by Ojiambo et al. (2017) to determine the stability of b based on the isotropic spread of CDM in the eastern United States. For both plant and animal diseases, b is approximately 2 over various spatial scales (Ojiambo et al., 2017). Ojiambo et al. (2017) examined if b varied from 2 and how much b varied over multiple years for CDM epidemics (Ojiambo et al., 2017). In that study, it was assumed that all epidemics were first observed at the same s_o. The authors found that b was unstable over multiple years i.e., $1.61 < b < 4.6$, using a temporal and spatial regression model (Ojiambo et al., 2017).

Pseudoperonospora cubensis is an obligate pathogen that requires a living host to reproduce and survive during the off-season. This characteristic of the pathogen is thought to restrict overwintering under natural conditions to frost-free areas below approximately 30-degree latitude in the United States (Ojiambo and Holmes, 2011; Ojiambo et al., 2015). Thus, the probability of pathogen transmission and disease spread from infected to disease-free cucurbit fields at different scales is influenced by the connectivity of fields in time and space. The determinants of pathogen dispersal and corresponding covariate information can be formulated in a model that explicitly considers network structure to represent disease spread (Gent et al., 2019). Networks contain nodes and links, where nodes are the entities of interest (e.g., field or host), and links connect the nodes in various ways. The links represent the potential transmission routes from an infected node to a susceptible node. Network models have been used in many scientific and

sociological disciplines. For example, in landscape ecology, networks have been used to quantify landscape connectivity and identify possible routes for dispersing organisms (Fletcher et al., 2011; Lookingbill et al., 2010; Minor and Urban 2008; Urban et al., 2009). In plant pathology, network models have been constructed to describe the spread of diseases caused by aerially dispersed pathogens (Gent et al., 2019; Sanatkar et al., 2015; Sutrave et al., 2012). In all these examples, networks were classified as either static (no change) or dynamic (changing with time).

A network whose structure does not change is known as a static network. Static networks are important because they highlight the impact of network structure, primarily where links occur, on the stability and connectivity of the network. These networks are widely used across many disciplines (Ferrari et al., 2014). For example, these networks have been used in landscape ecology to represent interpatch connectivity as a function of distance (Ferrari et al., 2014). Similarly, in plant pathology, such networks have been used to represent farms/fields/counties connectivity as a function of distance and a pathogen's dispersal characteristics (Gent et al., 2019; Sanatkar et al., 2015; Sutrave et al., 2012). Thus, static networks can identify potential pathways of disease spread and provide answers on specific characteristics of the described network structure. Also, static networks provide simple identification of invasion-prone locations based on between-location distances for management purposes (Ferrari et al., 2014).

Contrarily, a dynamic network is a network whose structure changes with time, wherein the nature of the change and notation used for the timing depends on the nature of the data (Enright and Kao, 2018; Ferrari et al., 2014). The dynamic networks model changes to node contacts themselves over time and provides insights regarding how node connectivity evolves during an invasion process. A dynamic network at a particular time may have a subset of the total static network (Ferrari et al., 2014). These networks have provided information regarding aerially borne

LDD pathogens (Gent et al., 2019; Sanatkar et al., 2015; Sutrave et al., 2012). In the latter examples, nodes are county centroids, while links (with calculated link weights) are the potential transmission routes. Sutrave et al. (2012) and Sanatkar et al. (2015) constructed dynamic networks to describe the spread of soybean rust caused by the obligate pathogen *Phakopsora pachyrhizi* (Sanatkar et al., 2015; Sutrave et al., 2012). Gent et al. (2019) constructed a dynamic network model for hop downy mildew caused by *Podosphaera macularis* and identified locations that could be targeted for replanting with resistant cultivars over multiple years for effective disease management (Gent et al., 2019).

In the present study, we formulate static and dynamic networks to understand the spread of CDM using epidemic data reported in the eastern United States. First, static networks are constructed and used to identify the relative resilience of generated networks from 2008 to 2016 using various network properties. Secondly, dynamic networks are constructed from static networks by adding wind velocity to the static networks such that *P.cubensis* can move along certain links at some time and other links at other times depending on the wind direction.

2. Methods

2.1. Data source

Cucurbit downy mildew epidemics from 2008 to 2016 in the eastern United States were obtained from the CDM ipmPIPE forecasting system (http://cdm.ipmpipe.org). This system collects CDM occurrence data in cucurbit fields, applies predictive models to the occurrence data, and communicates disease risk output to users such as cucurbit growers, extension personnel, and crop consultants (Ojiambo et al., 2011). During the study period (2008 - 2016), the disease was reported from sentinel and non-sentinel plots. Sentinel plots were fixed locations (measuring 50 ft

× 200 ft in size) placed within specific states to allow for disease monitoring every one to two weeks. All sentinel plots were georeferenced using the Global Positioning System. The cucurbits that were grown in the sentinel plots were *Cucumis sativus* (cucumber cv. Straight 8 and Poinsett 76), *Cucumis melo* (cantaloupe cv. Hales Best Jumbo), *Cucurbita pepo* (acorn squash cv. Table Ace), *Cucurbita maxima* (giant pumpkin cv. Big Max), *Cucurbita moschata* (butternut squash cv. Waltham), and *Citrullus lanatus* (watermelon cv. Micky Lee) (Ojiambo et al., 2011). The non-sentinel plots were not designated for regular surveillance. These included commercial fields, research plots, and home gardens. Since this data was available, it was included in the analysis to capture information where sentinel plots were not located. Typically, sentinel plots were planted earlier than regular cucurbit fields to allow for early disease detection. The collected dataset consisted of the date of first symptoms, month, affected host type, planting type, disease incidence, state, county, and disease location.

Data on hourly wind speed and direction for each location were derived from weather observations in the National Oceanic and Atmospheric Administration Integrated Surface Database (Smith et al., 2011) and provided by BASF (Research Triangle Park, Raleigh, NC) (more details on wind conversions are provided in chapter 3, section 2.1).

2.2. Static network analysis

Static networks were constructed for each epidemic year (2008 to 2016) using different *b* values. Both sentinel and non-sentinel plots were considered as nodes (*n*) in this work: 2008 (*n* = 154), 2009 (*n* = 220), 2010 (*n* = 156), 2011 (*n* = 127), 2012 (*n* = 173), 2013 (*n* = 204), 2014 (*n* = 114), 2015 (*n* = 215), and 2016 (*n* = 125). The links between nodes were created as a function of the between-node Euclidean distance and represent the potential transmission routes. Link weights

were also calculated for each link to represent the probability of transmission from an infected node I to a susceptible node S. An inverse power law dispersal kernel $y = (s_{ij})^{-b}$ was used to generate the links (details are provided in section 2.3). This inverse power-law model is the solution to Equation 1 above where s_{ij} is the distance between node *i* and node *j*, *b* is the spread parameter (Ojiambo et al., 2017), and *y* is the probability of transmission from node *i* to node *j* (Andersen et al., 2019). Values for parameter *b* tested here ranged from 1.51 to 3.36, and these were obtained from Ojiambo et al. (2017) based on their work on the isotropic spread of CDM in the eastern United States from 2008 to 2016. The between-node distance was calculated using the Haversine formula (Sinnot, 1984)

$$\begin{cases} q = sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos(\varphi_1)cos(\varphi_2)\, sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right) \\ d = R \times 2 \times atan2\left(\sqrt{q}, \sqrt{1-q}\right) \end{cases} \quad (3)$$

and x and y displacement vectors between two nodes were calculated using equirectangular projection

$$\begin{cases} x = R \times (\lambda_2 - \lambda_1)\, cos\left(\frac{\varphi_2 + \varphi_1}{2}\right) \\ y = R \times (\varphi_2 - \varphi_1) \end{cases} \quad (4)$$

where φ = latitude (radians), λ = longitude (radians), R = radius of the earth (mean = 6,371 km), and d = haversine distance between two nodes (node *i* = node 1 and node *j* = node 2).

2.2.1. Network connectivity threshold

To determine the potential links in an adjacency matrix ***A***, a threshold value (representing the probability of disease spread) was set above which, and below which, nodes were considered connected or unconnected, respectively. Thus, for a threshold value τ, the adjacency matrix ***A*** was populated with 1 if $y > \tau$ and 0 otherwise. The resulting link patterns form the basis of a pairwise

adjacency matrix $\boldsymbol{A}$. This matrix is constant in time with binary entries (0/1; not connected/connected). A network was created from the adjacency matrix i.e., the network of interest $G = (N, L)$ has node set N and link set L defined by the rule a link exists if and only if $y > \tau$ for every pair of nodes $(i, j) \in N$.

A range of arbitrary threshold values (ranging from 1×10^{-19} to 1×10^{-8}) was selected but bounded by values that produced a sparse network and a complete network. Different ranges of thresholds were also selected, with a constant incremental value, for each epidemic year. For example, the threshold range for the 2008 networks was 1×10^{-10} to 1×10^{-8} with a constant increment value of 5.2×10^{-10}. Three criteria were used to identify and select a threshold that was used to select a network in each year: i) the proportion of nodes in the giant component GC (part of a network that contains the majority of connected nodes), ii) no change in node ranking based on degree centrality, and iii) the extent of the connectedness of the generated network, a balance between a highly versus a sparsely connected network (Ames et al., 2011). The generated networks served as inputs for calculating network properties for each epidemic year as described below. All the networks were generated using the R programming language (R Core Development Team).

2.2.2. Static network properties

Network properties - for the static networks constructed above - were calculated using the *igraph* R package (Csardi and Nepusz, 2006). This study focused on seven properties of real networks (defined and explained in Table 2.1) based on their established and known relevance in models describing disease spread and pathogen transmission. These network properties are community, degree centrality, exponent of the degree distribution, density, diameter, giant

component, and the average shortest path. Additional information and relevance of these network properties in disease epidemiology are provided elsewhere (e.g., Danon et al., 2010).

2.3. Dynamic network construction

We added wind velocity to the static networks to create dynamic networks (more information on the wind data conversions is provided in Chapter 3, section 2.1). We modeled the SI dynamics as defined by Sutrave et al. (2012) that combines the static (constant during the cropping season) and the dynamic (vary during the cropping season) components. For discrete weekly time steps t ∈ {1, 2…T}, we have

$$\begin{cases} \alpha_{ij} = (s_{ij})^{-b} \\ \beta_{ij} = \frac{\vec{s_{ij}}.\vec{w_t}}{|\vec{s_{ij}}|} \\ u_{ij} = \alpha_{ij} \times \beta_{ij} \end{cases} \quad (5)$$

where α_{ij} is a function of the between-node distance s_{ij}, b is the spread parameter (Ojiambo et al., 2017), β_{ij} is the wind-based infection rate which is updated at each time step t, $\vec{s_{ij}}$ is the displacement vector between node i and node j, $\vec{w_t}$ is the wind vector at time t at node i, and u_{ij} is the link weight between node i and node j based on distance and wind direction at time t. The time-dependent u matrices were treated as weighted asymmetrical adjacency matrices. These weighted adjacency matrices were plotted at each time step to visualize the evolving dynamics on the static network. The in-degree and out-degree of the directed networks were calculated at each time step. Here, we define the out-degree as source strength such that a node is considered a source if out-degree > in-degree; otherwise, the node is a sink.

2.3.1. Error quantification in dynamic network

Two probabilities were first calculated at each time step *t*. The probability ϑ_i of node *i* not receiving infection from its neighbors,

$$\vartheta_i(t) = \prod_{j \in N_i}(1 - u_{ij}p_j(t)) \tag{6}$$

and the probability p_i of node *i* being infected at time *t*

$$p_i(t) = 1 - \big(1 - p_i(t-1)\big)\vartheta_i(t) \tag{7}$$

where p_j is the probability of node *j* being infected at time *t*, $u_{ij} \in [0,1]$ is the link weight, and N_i are node *i*'s neighbors (Sutrave et al., 2012). The nodes that were observed as infected were assigned value 1 and the healthy nodes were assigned value 0 at each time step. The error was then calculated as described by Sutrave et al. (2012):

i) $$\hat{E}_{in}(t) = \frac{\sum_{i=1}^{N_{in}(t)}(1-p_i(t))}{N_{in}(t)} \tag{8}$$

ii) $$\hat{E}_{hn}(t) = \frac{\sum_{i=1}^{N_{hn}(t)} p_i(t)}{N_{hn}(t)} \tag{9}$$

iii) $$\hat{E}(t) = \alpha\hat{E}_{in}(t) + (1-\alpha)\hat{E}_{hn}(t) \tag{10}$$

where $\hat{E}_{in}(t)$ is the mean error of the infected nodes at time step *t*, $N_{in}(t)$ is the total number of infected nodes at time step *t*, $\hat{E}_{hn}(t)$ is the mean error of the healthy nodes at time step *t,* $N_{hn}(t)$ is the total number of healthy nodes at time step *t*, $\hat{E}(t)$ is the total error, α is a weighting factor, and $p_i(t)$ is defined above. In this study, α = 0.8, i.e., the observed-infected nodes were given four times more weight than the observed-healthy nodes in evaluating the final error (Sutrave et al., 2012).

3. Results

3.1. Observed dynamics of disease spread

Across all the years, CDM was first observed in southern Florida in Miami-Dade County (Figure 2.1). The earliest date of disease occurrence in southern Florida was in February 2008, with most of the first cases of disease occurrence in other years being reported in February and March. Subsequently, detection of new disease occurrences progressed northward with time, with these new occurrences being reported later in more northern states, compared to occurrences in the southern states (Figure 2.1). The only exception to this pattern was observed in 2009, where an occurrence in southwestern Texas along the Gulf of Mexico was reported around the same time as disease occurrence reported in northern Florida and southern Georgia (Figure 2.1). The first occurrence of new disease cases in northern states (e.g., Michigan, New York, or Wisconsin) occurred considerably later than corresponding cases of first disease occurrence in southern states (e.g., Alabama, Georgia, or South Carolina) (Figure 2.1).

Across all epidemic years, the last set of new disease occurrences was reported in July, August, and September. The total number of states where CDM was reported ranged from 21 (in 2008 and 2016) to 25 (in 2013). The corresponding number of counties ranged from 85 to 178 across all epidemic years. In general, there appeared to be a spatial association between where CDM first occurred and the disease later developed within the region. Across all years, the maximum distance, a measure of epidemic extent, ranged from 2,491 km during the epidemic in 2012 to 3070.842 km during the epidemic in 2015.

3.2. Static networks

3.2.1 Network connectivity threshold

The connectivity threshold for generating networks was selected based on a combination of the proportion of nodes in the giant component (GC), no change in node ranking based on degree centrality, and the extent of the connectedness of the generated network. All the nodes were present in the GC at very low threshold values (e.g., 1×10^{-14} in 2016), with the nodes in the GC decreasing with increasing threshold values (Figure 2.2). Given that different values of b were used to create the networks for each year ($1.51 < b < 3.36$), the threshold value at which there was a drastic drop in the proportion of nodes varied across epidemic years. For example, in 2014, that threshold was around 1×10^{-16}, while that value was as high as 1×10^{-8} in 2009 (Figure 2.2). Nineteen nodes with the highest degree centrality scores (ranked in decreasing order for networks created with varying thresholds) further aided in selecting the connectivity threshold. For example, in 2008, at a threshold of 2.18×10^{-8}, all nodes ranked with high degree centrality scores were present in the network with a GC = 0.916 (Table 2.2). At a threshold of 2.71×10^{-8} with a GC = 0.76, two nodes (109 and 107) were unique for this threshold. These unique nodes were not present in other ranking lists for networks created by other thresholds. The nodes with the highest degree centrality scores identified across epidemic years were located in Connecticut, Massachusetts, Maryland, Michigan, North Carolina, Ohio, and Pennsylvania.

The networks were more connected at a threshold of 2.18×10^{-8} than at 2.71×10^{-8}, with the remaining thresholds having highly connected networks (Figure 2.3). A 2.18×10^{-8} threshold was thus selected for generating the 2008 network to achieve the desired balance in network connectivity. Similarly, in 2009, all nodes ranked with high degree centrality scores were present in the network at a threshold value of 7.83×10^{-9} with GC = 0.986 and 1.12×10^{-8} with GC = 0.95

(Table 2.3). However, nodes were not as well connected at $\tau = 1.12 \times 10^{-8}$ than at $\tau = 7.83 \times 10^{-9}$. The remaining thresholds resulted in either highly or sparsely connected networks (Figure 2.3). Thus, $\tau = 7.83 \times 10^{-9}$ was selected for generating the network in 2009. Using this same approach (i.e., the proportion of nodes in the GC, no change in node ranking and the extent of the connectedness of the generated network), threshold values selected to generate networks in 2010, 2011, 2012, 2013, 2014, 2015, and 2016 (see supplementary data) were 1.12×10^{-18}, 4.72×10^{-13}, 3.79×10^{-13}, 4.41×10^{-14}, 3.79×10^{-17}, 7.83×10^{-12}, and 3.37×10^{-12}, respectively. The thresholds differ because of different b values ($1.51 < b < 3.36$), different network sizes, and spatial locations.

3.2.2. Static network properties

The choice of threshold impacted the network properties. The paragraph below gives a description of the results rather than a comparison between years. The properties varied among epidemic years examined, with the extent of differences between epidemic years depending on individual properties (Table 2.4). For example, the number of neighbors for a node (i.e., average degree) ranged from 10 in 2014 to 24 in 2011. The exponent of degree distribution, i.e., the potential that a node would become an infected node and result in subsequent infections in neighboring nodes, ranged from 1.54 to 2.29 and was about 50% higher in 2014 than 2015 (Table 2.4). The static networks in 2014 had the highest number of groups of nodes with a high density of links within them than between groups (i.e., community). The number of communities in 2014 was about seven times higher in 2014 than in all other epidemic years except in 2010 and 2015 (Table 2.4). In 2010, the maximum shortest distance between the two most distant nodes (i.e., diameter) was 19 (Table 2.4), with diameters being the least in 2011 and 2016. The 2011 network

had a compact network with the shortest path of 2.29, while the network generated in 2010 was the least compact with the shortest path of 6.60 (Table 2.4).

Due to the choice of threshold, the size of the giant component was generally large (GC ≥ 90%), indicating highly connected networks across epidemic years except in 2010, where GC was relatively lower with a single component containing 89% of the nodes in the network (Table 2.4). The highest density of links was observed in 2011 with a value of 0.20, which was about twice as high as those observed in other epidemic years except in 2015. Degree centrality (or simply degree) was highest in 2011 with a value of 38 and lowest in 2014 with a value of 20, with centrality values in the remaining epidemic years ranging from 24 to 32 (Table 2.4). Thus, the number of nodes that could potentially serve as 'superspreaders' of CDM was highest in 2011 and lowest in 2014. Most of these potentially superspreader nodes were located in Michigan, Virginia, Ohio, North Carolina, South Carolina, and Maryland.

3.2.3. Dynamic network model for cucurbit downy mildew

The dynamic network model for CDM revealed an emerging and evolving network and provided information that was not captured from the static network representation of disease spread (Figure 2.5). Early on, nodes far away from the source were not infected in all epidemic years, so the transmission did not occur to their neighbors, and the link probabilities were low. However, the links between nodes closest to the initial point infection (open square) in southern Florida had the highest link probabilities early in the season (i.e., week 10), while the link probabilities between nodes were low elsewhere in the network (Figure 2.5). As the disease spread proceeded in time and space, link probabilities increased for nodes that were more distant from the node where the initial occurrence was observed, although probabilities remained low for other

nodes (Figure 2.5). For example, in 2013, links between nodes in Georgia closest to the initial disease outbreak in southern Florida had the highest link probabilities at week = 10. At week 15, with the advance in the epidemic, the link probabilities between these nodes and those in South Carolina increased, while nodes elsewhere in northern states remain low. At week 20, link probabilities between nodes in South Carolina and North Carolina increased, while link probabilities were low between nodes in more northern states or in states where no disease was present. New links with moderate levels of probability were emerging between nodes in North Carolina, Virginia, and West Virginia during this time. At week 25, link probabilities increased between nodes in North Carolina, Virginia, New York, Pennsylvania, Ohio, and Michigan (Figure 2.5). A similar emerging and evolving pattern of the network was observed in other epidemic years except that link probabilities between nodes in different states and the strength of these probabilities differed between years (Figure 2.5).

In all epidemic years, the out-degrees of the directed networks at each time step (source strength) from the dynamic network models were always less than the in-degree (Figure 2.6) by week 35. Most of the nodes in the network tended to act as sinks (i.e., zero or low source strength) because of overlapping sets of neighbors. However, the location of these sink nodes varied among epidemic years. For example, in 2008 and 2010, most of the nodes that tended to act as sinks were located in the Midwest and the northeast region, while in 2013, these nodes were primarily located in the Atlantic coast region (Figure 2.6). In addition, out- and in-degrees were comparatively lower in 2014 than in other epidemic years. Often, the strongest source nodes were found in areas with many neighboring nodes. However, some nodes acted as sources mainly when they were near the initial source of disease outbreaks in South Florida, especially in 2008, 2013, and 2014 (Figure 2.6).

3.2.4. Error quantification in dynamic networks

The mean absolute errors varied depending on the epidemic year and time step used to construct the dynamic networks. The mean errors were calculated at weekly time steps and averaged monthly. Across all time steps, mean absolute errors ranged from 0.099 in 2016 to 0.353 in 2010 (Table 2.5). Epidemic years with comparable low errors as reported in 2016 were 2014 and 2015 that had absolute errors of 0.168 and 0.121, respectively. Overall, the mean absolute errors were lowest at the start of the epidemic during the January to February transition (error = 0.002) and February to March transition (error = 0.035). The errors increased steadily after that and were highest at the end of the epidemic of each year during the June to July transition (error = 0.343) and July to August transition (error = 0.332) (Table 2.5). The mean absolute error across all time steps from 2008 to 2016 was 0.194.

4. Discussion

In this study, models that explicitly consider network structures to represent disease spread were formulated to describe the dispersal of *Pseudoperonospora cubensis* and spread of cucurbit downy mildew from infected to disease-free sentinel and non-sentinel plots in the eastern United States. Although sentinel plot data is more reliable than non-sentinel plots, the non-sentinel plots data was included because the data was available. Also, the non-sentinel plots data accounted for areas where sentinel plots were not located in the eastern U.S. In addition, reports from commercial fields, home gardens, and research plots are valuable locations at which infection became established and may have likely played a role in pathogen transmission and disease spread.

Based on a static network model, several network properties that may be useful predictors of risk of infection and time to infection during disease occurrence in cucurbit fields were

identified. These identified properties may have important implications for monitoring and controlling CDM within the region. Further, a dynamic network model was developed to model the CDM spread in the eastern United States. The dynamic model identified locations that contributed most to disease spread across the landscape and showed that source strength (out-degree) was always less than the in-degree, with most nodes in the network tending to act as sinks. The strongest source nodes tended to be near many nodes, while some nodes acted as sources mainly if they were near the initial focus. The dynamic model performed well with low errors, especially early in the growing season across the epidemic years examined.

The characteristics of a disease-spreading process are determined by the topology of the network (Newman, 2002). In this study, degree centrality identified highly connected nodes in the network that were responsible for disease spread in specific CDM epidemics. Across all epidemic years, the more central nodes were primarily located in Michigan in the Great Lakes region, Ohio in the mid-west, Maryland, North Carolina, South Carolina, and Virginia along the Atlantic coast. Thus, these nodes could be reasonable targets for more intensive sampling, monitoring, and management to reduce inoculum production that drives infection in neighboring cucurbit fields. Degree centrality is a valuable measure for identifying important nodes in static networks of several pathosystems to inform strategic and tactical decisions about disease management in plant and animal systems (Gent et al., 2019; Christley et al., 2005; Kiss et al., 2006; Xing et al., 2020). Other centrality measures such as eigenvector, betweenness, and closeness that help identify central nodes in networks (Wang 2003; Christley et al., 2005; Xing et al., 2020) were not investigated in this study. It would be useful to assess the consistency of the results reported here using these measures.

The ability to accurately estimate the threshold for disease spread is critical as it allows for predictions of epidemic development and the identification of appropriate containment measures. Connectivity networks are often constructed by applying a threshold to the resulting network to retain only the most epidemiologically significant relationships between nodes connected by links. This threshold application step is critical as it can introduce errors in the network construction and inform of both false negatives and false positives (Perkins et al., 2009). There is no consensus on what threshold method can be used due to the inherent differences across pathosystems. Gent et al. (2019) applied a quantitative approach that selected thresholds based on a receiver operating characteristic curve analysis that maximized binary classification to classification accuracy in network analysis of hop powdery mildew. There are other ways of creating networks besides the thresholding approach used in this study (e.g., Andersen et al., 2019).

In this study, highly connected networks that were too large to make exciting inferences were generated at low thresholds, while sparsely connected networks were generated with high threshold values. Both the highly connected networks and the sparsely connected networks overshadow the effects of network properties such as path length, density, and diameter on network structure (Ames et al., 2011). Thus, thresholds were selected to generate intermediate density networks that maximized the impact of disease dynamics on the properties of the network structure in each epidemic year. Networks of intermediate density have also been reported to profoundly impact disease behavior (Ames et al., 2011).

The dynamic model developed in this study utilized the inverse power-law function to model (i.e., generate link weights based on wind speed and direction) disease spread from infected to disease-free fields some distance away from the source. The power-law function is useful in modeling diseases caused by aerially borne plant pathogens (Ferrandino 1993; Madden et al.,

2007; Mundt et al., 2009). Errors associated with the power-law function, cut-off, and *b* value, were very low at the start of epidemic onset but steadily increased as new disease occurrences were reported. This increase in prediction error with increased disease outbreaks may partly be explained by substantial 'hops' in disease outbreaks during the epidemic period (Sutrave et al., 2012). Substantial 'hops' during the epidemic period are not unexpected for *P. cubensis*, whose inoculum can be estimated to be transported over 1,000 km (Ojiambo and Holmes, 2011).

In other pathosystems, the negative exponential and gravity models are useful in modeling disease spread. For example, gravity models were used to model the soybean rust in the United States (Sutrave et al., 2012; Xing et al., 2020). The gravity model used by Sutrave et al. (2012) resulted in low prediction errors, a characteristic that was attributed to the absence of large uninfected regions dividing infected areas. The choice of a function to model dispersal events (or disease spread) will depend on the product of the potential amount of inoculum produced at the source and potential level of host availability at the sink, two attributes that likely vary among different pathosystems. The prediction ability of the dynamic model used in this study could be improved by incorporating effects of weather factors such as temperature, rainfall, and UV radiation (Kanetis et al., 2010; Neufeld et al., 2018) or accounting for the source strength at a given location (Neufeld et al., 2013; Ojiambo et al., 2015).

In this study, the dynamic network revealed an emerging and evolving network that provided useful information that could inform decision-making to manage cucurbit downy mildew in the eastern United States. By evaluating the emergent network as a series of time-dependent adjacency matrices, the dynamic model identified areas that may act as sources and promote CDM spread. While most of the nodes in the network tended to act as sinks, these nodes were located in the Midwest, Northeast, and Atlantic coast regions. The dynamic model also facilitated

visualization of the prominent time-dependent pathways of pathogen dispersal. One of the critical challenges in predicting disease spread has been to estimate the probability and timing of disease outbreaks in specific locations (Meentemeyer et al., 2011; Fitzpatrick et al., 2012) and determine where and when the introduction of inoculum can impact the extent of an epidemic. By varying initial conditions related to the directionality of wind and associated source strength of a cucurbit field, the dynamic network developed here can be used to determine nodes that result in the rapid spread and a large epidemic extent and thus pose the greatest risk of epidemic expansion across the region (Ferrari et al., 2014; Ojiambo et al., 2017).

In the United States, CDM expansion in space and time is a characteristic of a dispersive wave epidemic with accelerating velocity (Ojiambo et al., 2017). Thus, the first fungicide spray timing is critical in limiting the northward advance of CDM epidemics during the growing season (Neufeld et al., 2018). However, the timing of its spread from overwintering sources in southern Florida is the most uncertain feature within the prediction framework (Ojiambo and Holmes, 2011). The dynamic model developed in this study could be used as a valuable decision support system to inform uncertain situations concerning locations of initial disease outbreaks. Such an approach will need to be complemented with scouting efforts for disease outbreaks and pathogen detection to provide knowledge of inoculum sources and pathogen dispersal across fields. This study's application of connectivity analysis assumed homogeneity in the host response to *P. cubensis* and favorable weather for infection and spread. However, the environmental factors are likely to vary across locations, and the dynamic model developed here can further be improved by incorporating prevailing weather factors in the dispersal framework of the pathogen (Margosian et al., 2009).

References

1. Ames, G. M., George, D. B., Hampson, C. P., Kanarek, A. R., McBee, C. D., et al. 2011. Using network properties to predict disease dynamics on human contact networks. Proc. R. Soc. B 278:3544-3550.
2. Andersen, K. F., Buddenhagen, C. E., Rachkara, P., Gibson, R., Kalule, S., Phillips, D., and Garrett, K. A. 2019. Modeling epidemics in seed systems and landscapes to guide management strategies: The case of sweet potato in northern Uganda. Phytopathology 109: 1519-1532.
3. Campbell, C. L., and Madden, L. V. 1990. Introduction to Plant Disease Epidemiology. Wiley, New York.
4. Christley, R. M., Pinchbeck, G. L., Bowers, R. G., Clancy, D., Frnech, N. P., Bennett, R., and Turner, J. 2005. Infection in social networks: Using network analysis to identify high-risk individuals. Am. J. Epidemiol. 162:1042-1031.
5. Csardi, G., and Nepusz, T. 2006. The igraph software package for complex network research. InterJournal, Complex Systems, 1695. http://igraph.org.
6. Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V, and Vernon, M. C. 2010. Networks and the epidemiology of infectious disease. Interdiscip. Perspect. Infect. Dis. Volume 2011, Article ID 284909.
7. Enright, J., and Kao, R. R. 2018. Epidemics on dynamic networks. Epidemics 24: 88-97.
8. Ferrari, J. R., Preisser, E. L., and Fitzpatrick, M. C. 2014. Modeling the spread of invasive species using dynamic network models. Biol. Invasions 16: 949-960.
9. Ferrandino, F. J. 1993. Dispersive epidemic waves: Focus expansion with a linear planting. Phytopathology 83: 795-802.

10. Fitzpatrick, M. C., Preisser, E. L., Porter, A., Elkinton, J., and Ellison, A. M. 2012. Modeling range dynamics in heterogeneous landscapes: invasion of the hemlock woolly adelgid in eastern North America. Ecol. Appl. 22:472-486.
11. Fletcher, R. J., Acevedo, M. A., Reichert, B. E., Pias, K. E., and Kitchens, W. M. 2011. Social network models predict movement and connectivity in ecological landscapes. Proc. Natl. Acad. Sci. 108:19282-19287.
12. Gent, D. H., Bhattacharyya, S., and Ruiz, T. 2019. Prediction of spread and regional development of hop powdery mildew: A network analysis. Phytopathology 109:1392-1403.
13. Hijmans, R. J. 2017. geosphere: Spherical Trigonometry. R Package Version 1.5-7. https://cran.r-project.org/web/packages/geosphere/index.html
14. Holmes, G. J., Ojiambo, P. S., Hausbeck, M. K., Quesada-Ocampo, L., and Keinath, A. P. 2015. Resurgence of cucurbit downy mildew in the United States: A watershed event for research and extension. Plant Dis. 99:428-441.
15. Kanetis, L., Holmes, G. J., and Ojiambo, P. S. 2010. Survival of *Pseudoperonospora cubensis* sporangia exposed to solar radiation. Plant Pathol. 59:313-323.
16. Kiss, I. Z., Green, D. M., Kao, R. R. 2006. The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread. J. R. Soc. Interface 3:669-677.
17. Lookingbill, T. R., Gardner, R. H., Ferrari, J. R., and Keller, C. E. 2010. Combining a dispersal model with network theory to assess habitat connectivity. Ecol. Appl. 20:427-441.

18. Madden, L. V., Hughes, G., and van den Bosch, F. 2007. The Study of Plant Disease Epidemics. APS Press, St. Paul, MN.

19. Margosian, M. L., Garrett, K. A., Hutchinson, J. M. S., and With, K. A. 2009. Connectivity of the American agricultural landscape: Assessing the national risk of crop pest and disease spread. Bioscience 59:141-151.

20. Meentemeyer, R. K., Cunniffe, N. J., Cook, A. R., Filipe, J. A. N., Hunter, R. D., Rizzo, D. M., and Gilligan, C. A. 2011. Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990–2030). Ecosphere 2:art 17. doi:10.1890/ES10-00192.1

21. Minor, E. S., Urban, D. L. 2008. A graph-theory framework for evaluating landscape connectivity and conservation planning. Conserv. Biol. 22:297-307.

22. Mundt, C. C., Sackett, K. E., Wallace, L. D., Cowger, C., and Dudley, J. P. 2009a. Long-distance dispersal and accelerating waves of disease: empirical relationships. Am. Nat. 173:456-466.

23. Mundt, C. C., Sackett, K. E., Wallace, L. D., Cowger, C., and Dudley, J. P. 2009b. Aerial dispersal and multiple scale spread of epidemic disease. EcoHealth 6:546-552.

24. Nathan, R., Perry, G., Cronin, J., Strand, A., & Cain, M. (2003). Methods for estimating long-distance dispersal. *Oikos, 103*, 261-273.

25. Newman, M. E. J. 2002. Spread of epidemic disease on networks. Phys. Rev. E 66: 016128.

26. Neufeld, K. N., Isard, S. A., and Ojiambo, P. S. 2013. Relationship between disease severity and escape of *Pseudoperonospora cubensis* sporangia from a cucumber canopy during downy mildew epidemics. Plant Pathol. 62:1366-1377.

27. Neufeld, K. N., Keinath, A. P., Gugino, B. K., McGrath, M. T., Sikora, E. J., Miller, S. A., Ivey, M. L., Langston, D. B., Dutta, B. Keever, T., Sims, A., and Ojiambo, P. S. 2018. Predicting the risk of cucurbit downy mildew in the eastern United States using an integrated aerobiological model. Int. J. Biometerol. 62:655-668.

28. Ojiambo, P. S., and Holmes, G. J. 2011. Spatiotemporal spread of cucurbit downy mildew in the eastern United States. Phytopathology 101: 451-461.

29. Ojiambo, P. S., Holmes, G. J., Britton, W., Keever, T., Adams, M. L., Babadoost, M., Bost, S. C., Boyles, R., Brooks, M., Damicone, J., Draper, M. A., Egel, D. S., Everts, K. L., Ferrin, D. M., Gevens, A. J., Gugino, B. K., Hausbeck, M. K., Ingram, D. M., Isakeit, T., Keinath, A. P., Koike, S. T., Langston, D., McGrath, M. T., Miller, S. A., Mulrooney, R., Rideout, S., Roddy, E., Seebold, K. W., Sikora, E. J., Thornton, A., Wick, R. L., Wyenandt, C. A., and Zhang, S. 2011. Cucurbit downy mildew ipmPIPE: a next generation web-based interactive tool for disease management and extension outreach. Online. Plant Health Progress. Online publication. doi:10.1094/PHP-2011-0411-01-RV.

30. Ojiambo, P. S., Gent, D. H., Mehra, L. K., Christie, D., and Magarey, R. 2017. Focus expansion and stability of the spread parameter estimate of the power law model for dispersal gradients. PeerJ 5:e3465.

31. Ojiambo, P. S., Gent, D. H., Quesada-Ocampo, L. M., Hausbeck, M. K., and Holmes, G. J. 2015. Epidemiology and population biology of *Pseudoperonospora cubensis*: a model system for management of downy mildews. Annu. Rev. Phytopathol. 53: 223-246.

32. Perkins, Andy & Langston, Michael. (2009). Threshold Selection in Gene Co-Expression Networks Using Spectral Graph Theory Techniques. BMC bioinformatics. 10 Suppl 11. S4. 10.1186/1471-2105-10-S11-S4.

33. R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
34. Sanatkar, M. R., Scoglio, C., Natarajan, B., Isard, S. A., and Garrett, K. A. 2015. History, epidemic evolution, and model burn-in for a network of annual invasion: Soybean rust. Phytopathology 105:947-955.
35. Severns, P. M., Sackett, K. E., Farber, D. H., and Mundt, C. C. 2019. Consequences of long-distance dispersal for epidemic spread: patterns, scaling, and mitigation. Plant Dis. 103: 177-191.
36. Scherm, H. 1996. On the velocity of epidemic waves in model plant disease epidemics. Ecol. Model. 87:217-222.
37. Smith, A., Lott, N., and Vose, R. The integrated surface database: recent developments and partnerships. Bull. Am. Meteorol. Soc. 92: 704-708.
38. Sutrave, S., Scoglio, C., Isard, S., Hutchinson, J. M. S., and Garrett, K. A. 2012. Identifying highly connected counties compensates for resource limitations when evaluating national spread of an invasive pathogen. PLoS One 7:e37793.
39. Sinnott, R. W. 1984. Virtues of the Haversine. Sky Telescope 68:158.
40. Urban, D.L., Minor, E. S., Treml, E. A., and Schick, R. S. 2009. Graph models of habitat mosaics. Ecol. Lett. 12:260-273.
41. Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C. 2003. Epidemic spreading in real networks: An eigenvalue viewpoint. Pages 25-34 in: Proceedings of the 22nd International Symposium on Reliable Distributed Systems. Florence, Italy.

42. Xing, Y., Hernandez Nopsa, J. F., Andersen, K. F., Andrade-Piedra, J. L., Been, F. D., et al. 2020. Global cropland connectivity: A risk factor for invasion and saturation by emerging pathogens and pests. Bioscience 70:744-758.

Tables

Table 2.1. Network properties used to characterize the spread of cucurbit downy mildew in the eastern United States.

Property	Definition	Relevance in pathogen transmission and disease spread
Average shortest path	The mean shortest path between nodes considering all possible pairs of nodes in a network	A small value indicates a compact network. It is a measure of the efficiency of disease spread on a network
Diameter	The maximum shortest path between the any pair of nodes in the network	It measures how many intermediary nodes infection must travel to spread disease from any node to any other node
Density	The ratio of the present links in a network to the maximum possible number of links	A higher density of connections can enable greater pathogen transmission in the network
Community	Sets of nodes that have a high density of links within them and a lower density of links between groups	Interventions targeting nodes bridging communities are more effective in controlling disease spread
Giant component	A section of a network that contains the majority of nodes in the network	Influences the extent of disease propagation since there are no paths to nodes in other components; an introduction of infection will be able to reach all nodes in the giant component (Danon et al., 2010)
Degree distribution	The probabilities that a node chosen at random will have a given degree	Captures heterogeneity in the potential of a node to be infected and cause additional infections. Knowing the distribution is crucial in understanding disease spread dynamics
Degree centrality	The number of links a node in the network (for directed networks, in-degree is the number of incoming links and out-degree is the number of outgoing links)	The node degree of an epidemic starting point influences epidemic outcomes; those with a high degree may be 'superspreaders' once infected. The higher the degree of an infected node, the most likely it is to cause a large number of subsequent infections

Table 2.2. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2008.

	GC = 0.922		GC = 0.922		GC = 0.916		GC = 0.760		GC = 0.500		GC = 0.468	
Node rank[a]	$\tau = 1.14 \times 10^{-9}$		$\tau = 1.66 \times 10^{-9}$		$\tau = 2.18 \times 10^{-9}$		$\tau = 2.71 \times 10^{-9}$		$\tau = 4.27 \times 10^{-9}$		$\tau = 6.35 \times 10^{-9}$	
1	132 [b]	45 [c]	71	34	58	33	71	28	1	21	6	16
2	126	45	62	34	62	33	62	27	68	18	5	16
3	113	44	58	34	70	30	58	27	63	17	3	16
4	111	43	119	34	116	29	116	26	60	17	1	16
5	70	42	112	33	113	25	112	22	9	17	64	16
6	114	42	122	33	112	25	67	22	6	17	63	15
7	120	41	113	32	68	25	63	22	5	17	60	15
8	115	40	70	32	67	25	68	22	65	17	9	15
9	122	40	95	32	63	25	64	21	64	17	8	15
10	123	39	111	30	122	25	60	21	8	17	7	15
11	121	38	68	29	64	24	120	21	7	17	65	15
12	119	38	123	29	60	23	119	20	4	16	48	14
13	116	38	121	28	123	23	115	20	71	16	114	13
14	112	38	126	28	121	22	114	19	62	16	80	12
15	95	38	132	27	119	22	113	19	122	16	**78**	12
16	58	38	114	26	108	22	111	18	50	16	71	12
17	130	37	67	26	120	22	**109**	18	**49**	16	61	12
18	**104** [b]	36	64	26	115	21	108	18	48	16	55	12
19	71	36	63	26	111	21	**107**	18	**47**	15	50	12

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value. The GC value is calculated for each network. The same value for two networks does not mean that the networks are the same.

Table 2.3. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2009.

	GC = 1.00		GC = 0.991		GC = 0.986		GC = 0.950		GC = 0.842		GC = 0.367	
Node rank[a]	$\tau = 1.0 \times 10^{-9}$		$\tau = 4.41 \times 10^{-9}$		$\tau = 7.83 \times 10^{-9}$		$\tau = 1.12 \times 10^{-8}$		$\tau = 1.47 \times 10^{-8}$		$\tau = 1.81 \times 10^{-8}$	
1	**154** [b]	189[c]	211 [b]	50	79	35	93	28	93	23	91	20
2	**55**	188	201	50	74	35	82	28	109	22	89	20
3	**54**	188	**129**	49	201	33	88	27	92	22	75	20
4	**53**	188	128	49	159	33	91	26	91	22	109	19
5	**50**	188	**81**	49	109	33	89	26	89	22	83	19
6	**48**	188	**133**	48	93	33	83	26	83	22	76	19
7	**164**	181	**126**	48	82	33	76	26	82	22	95	18
8	**193**	179	**124**	48	212	32	75	26	76	22	92	18
9	213	177	**87**	48	136	32	136	25	75	22	90	18
10	**206**	177	212	47	118	32	92	25	90	21	88	18
11	152	177	208	47	114	32	90	25	88	21	93	17
12	**166**	176	**204**	47	90	32	79	25	125	19	82	17
13	**155**	176	136	47	76	32	127	24	**97**	19	80	17
14	163	175	**135**	47	208	31	109	24	95	19	79	17
15	**132**	175	118	47	130	31	95	24	79	19	128	16
16	**165**	174	213	46	127	31	158	23	74	19	125	16
17	156	174	160	46	**111**	31	130	23	128	18	119	16
18	160	173	159	46	92	31	125	23	121	18	117	16
19	157	173	158	46	91	31	121	23	119	18	114	16

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.

[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.

[c] The calculated degree centrality value.

Table 2.4. Measured properties of static networks for the spread of cucurbit downy mildew in the eastern United States based on disease epidemics reported from 2008 to 2016.

	Network property							
Year	Average degree	Exponent of the degree distribution[a]	Community number	Diameter	Density	Proportion of nodes in GC[b]	Average shortest path	Average degree [c]
2008	13.77	1.70	13	17	0.09	0.92	5.44	25
2009	16.54	1.63	13	15	0.08	0.98	5.15	32
2010	12.43	1.83	23	19	0.08	0.89	6.60	28
2011	24.93	2.05	11	9	0.20	0.94	2.92	38
2012	15.57	1.67	12	16	0.09	0.98	5.20	28
2013	14.75	1.62	14	14	0.07	0.96	4.89	27
2014	10.72	2.29	73	14	0.10	0.99	5.29	20
2015	12.61	1.54	22	17	0.06	0.90	5.90	24
2016	12.59	1.71	10	10	0.10	0.94	4.42	24

[a] Denotes the exponent of degree distribution, where the higher the value is for a node, the most likely that node will become infected and result in subsequent infections in neighboring nodes.

[b] GC denotes giant component, i.e., a single component that contains the majority of nodes in the network, where a high value depicts a tightly connected network.

[c] Number of links that a node has to other nodes in the network based on a mean of 19 nodes ranked as most important in each epidemic year.

Table 2.5. Absolute errors for different time steps for network models used to characterize the spread of cucurbit downy mildew in eastern United States.

	Epidemic year									
Time step	2008	2009	2010	2011	2012	2013	2014	2015	2016	**Mean[b]**
Jan-Feb	-[a]	-	-		0.003	-	-	0.001	-	**0.002**
Feb-Mar	0.006	-	-	0.008	0.003	0.157	0.037	0.001	-	**0.035**
Mar-Apr	0.095	0.003	0.093	0.124	0.184	0.310	0.078	0.180	0.014	**0.120**
Apr-May	0.379	0.232	0.343	0.347	0.350	0.290	0.136	0.181	0.022	**0.253**
May-Jun	0.317	0.296	0.411	0.395	0.462	0.163	0.180	0.134	0.089	**0.272**
Jun-Jul	0.343	0.533	0.414	0.554	0.522	0.156	0.218	0.186	0.165	**0.343**
Jul-Aug	0.249	0.441	0.504	0.409	-	-	0.358	0.161	0.203	**0.332**
Mean[c]	**0.231**	**0.301**	**0.353**	**0.306**	**0.254**	**0.215**	**0.168**	**0.121**	**0.099**	**0.194**

[a] No disease outbreaks occurred reported in this time step and thus no error was computed for this time step.
[b] Mean absolute error across all epidemic years at each time step.
[c] Mean absolute error across all time steps within an epidemic year.

Figures

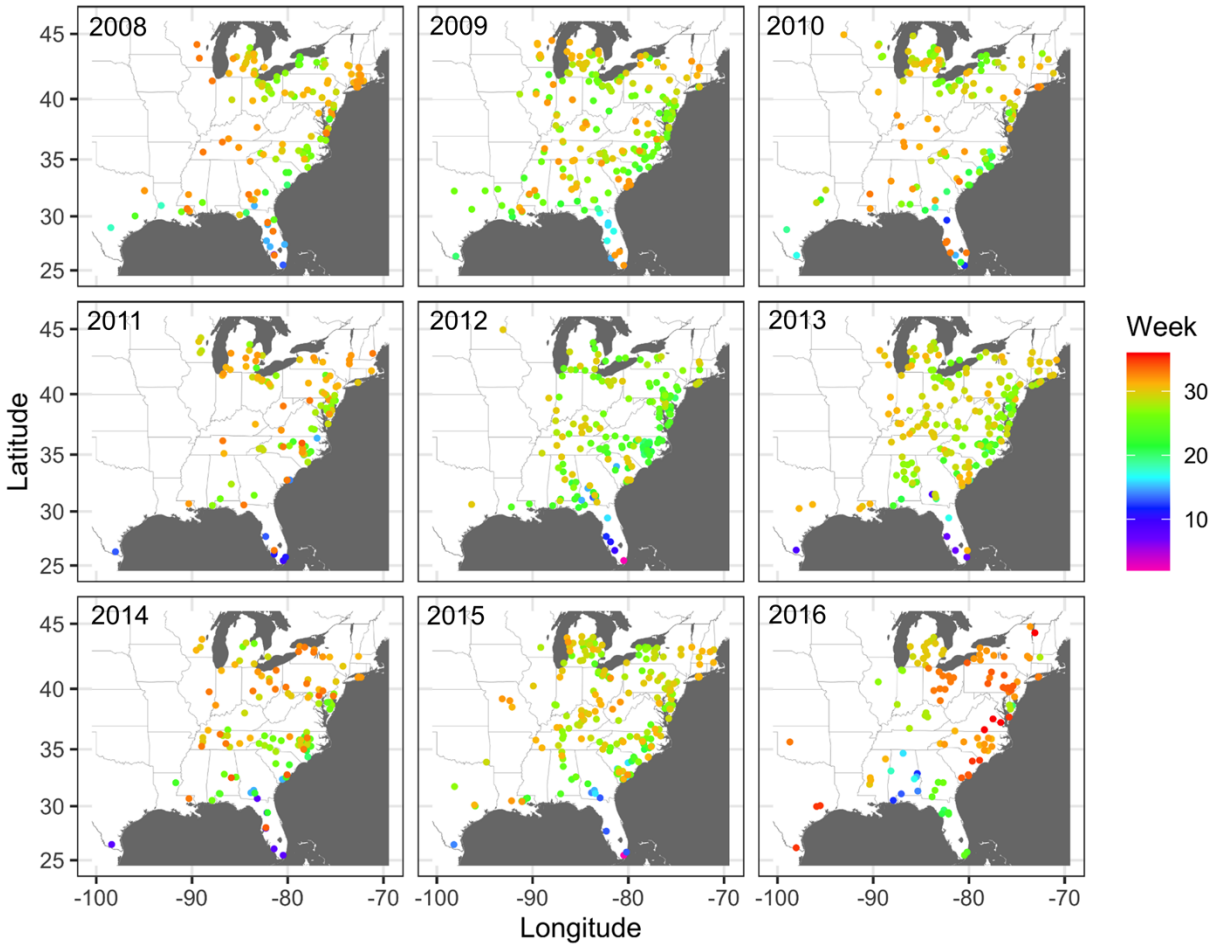


Figure 2.1. Locations of cucurbit downy mildew outbreaks in the eastern United States from 2008 to 2016. The locations are color-coded based on the week of the year. Every year, cucurbit downy mildew was reported early in Florida and along the Gulf of Mexico before the northern states.

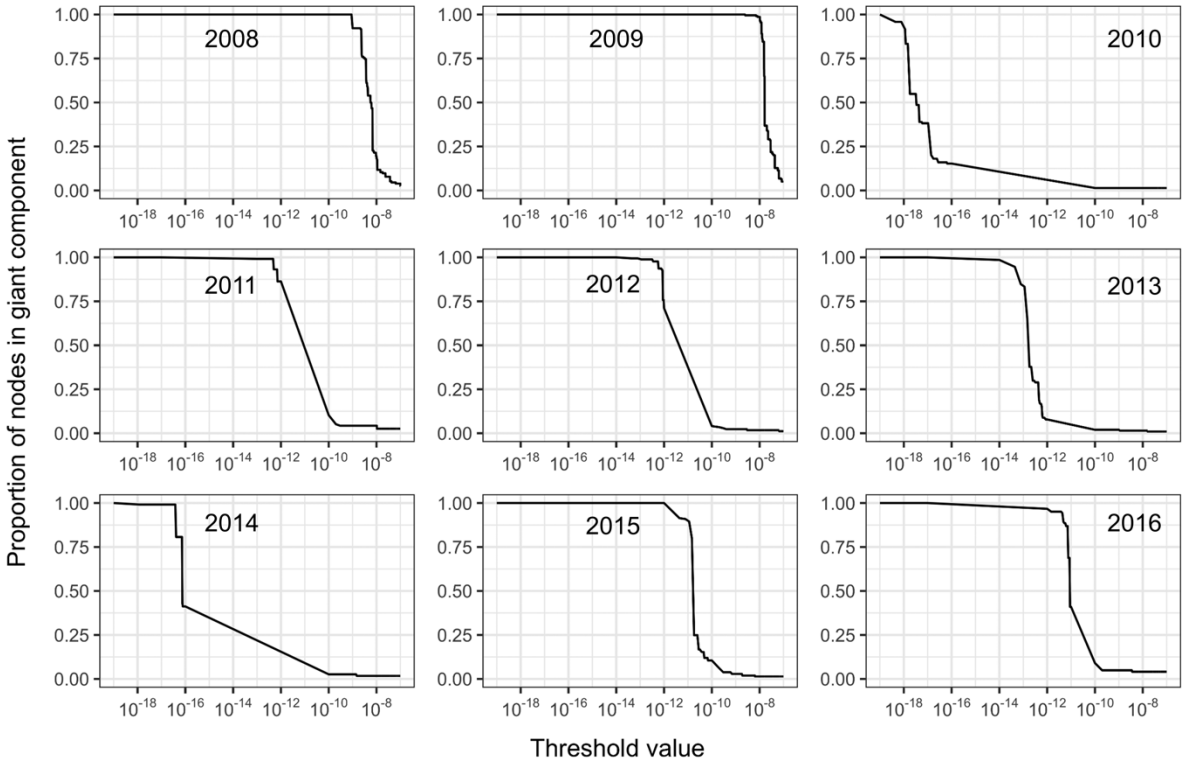


Figure 2.2. The proportion of nodes in the giant component is influenced by the connectivity threshold for cucurbit downy mildew epidemics from 2008 to 2016 in the eastern United States.

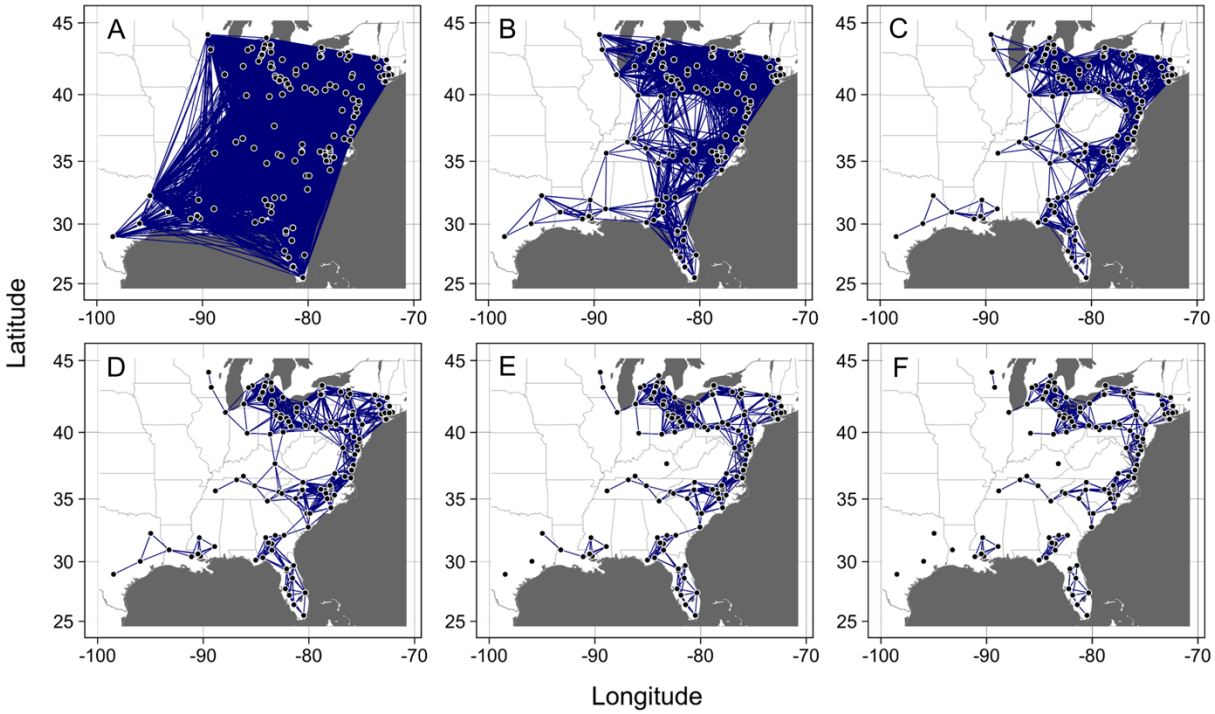


Figure 2.3. Examples of 2008 networks generated with spread parameter *b* = 1.61. The thresholds for the networks shown in panels A - F are 1.00×10^{-10}, $6.21 \text{ X } 10^{-10}$, 1.14×10^{-9}, 1.66×10^{-9}, $2.18 \text{ X } 10^{-9}$, and 2.71×10^{-9}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

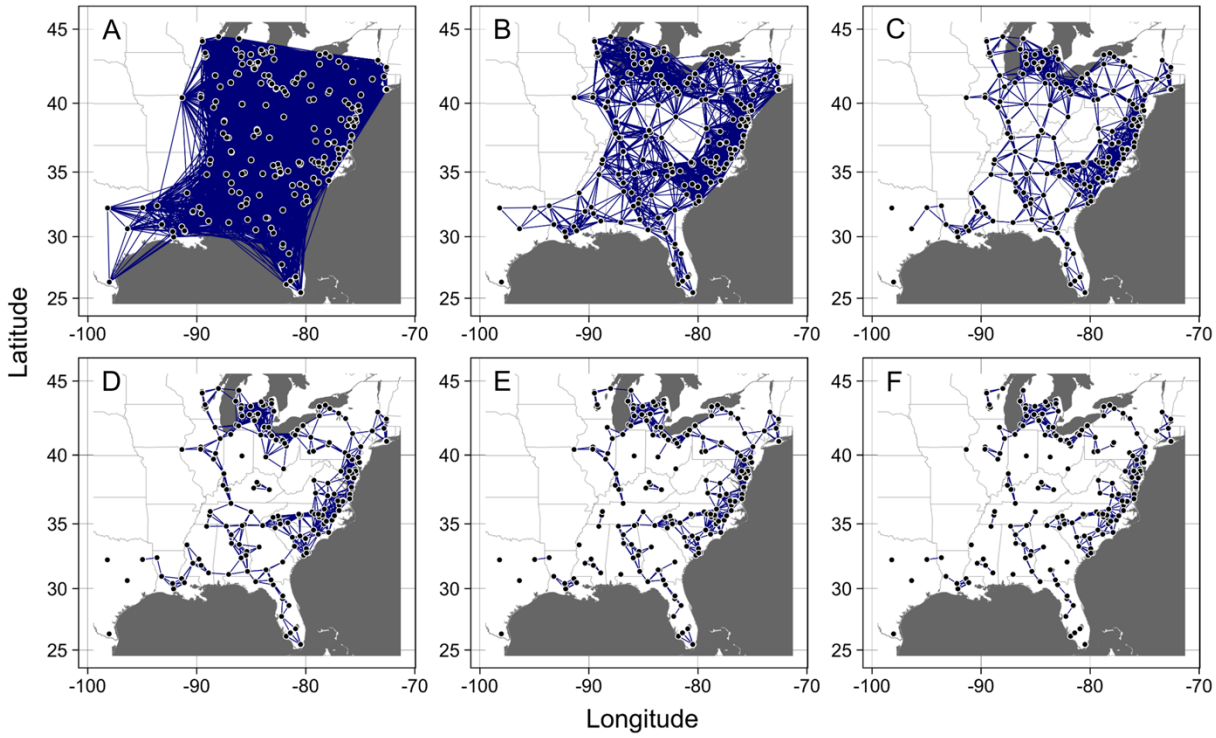


Figure 2.4. Examples of 2009 networks generated with spread parameter *b* = 1.51. The thresholds for the networks shown in panels A - F are 1.00×10^{-9}, 4.41×10^{-9}, 7.83×10^{-9}, 1.12×10^{-8}, 1.47×10^{-8}, and 1.81×10^{-8}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

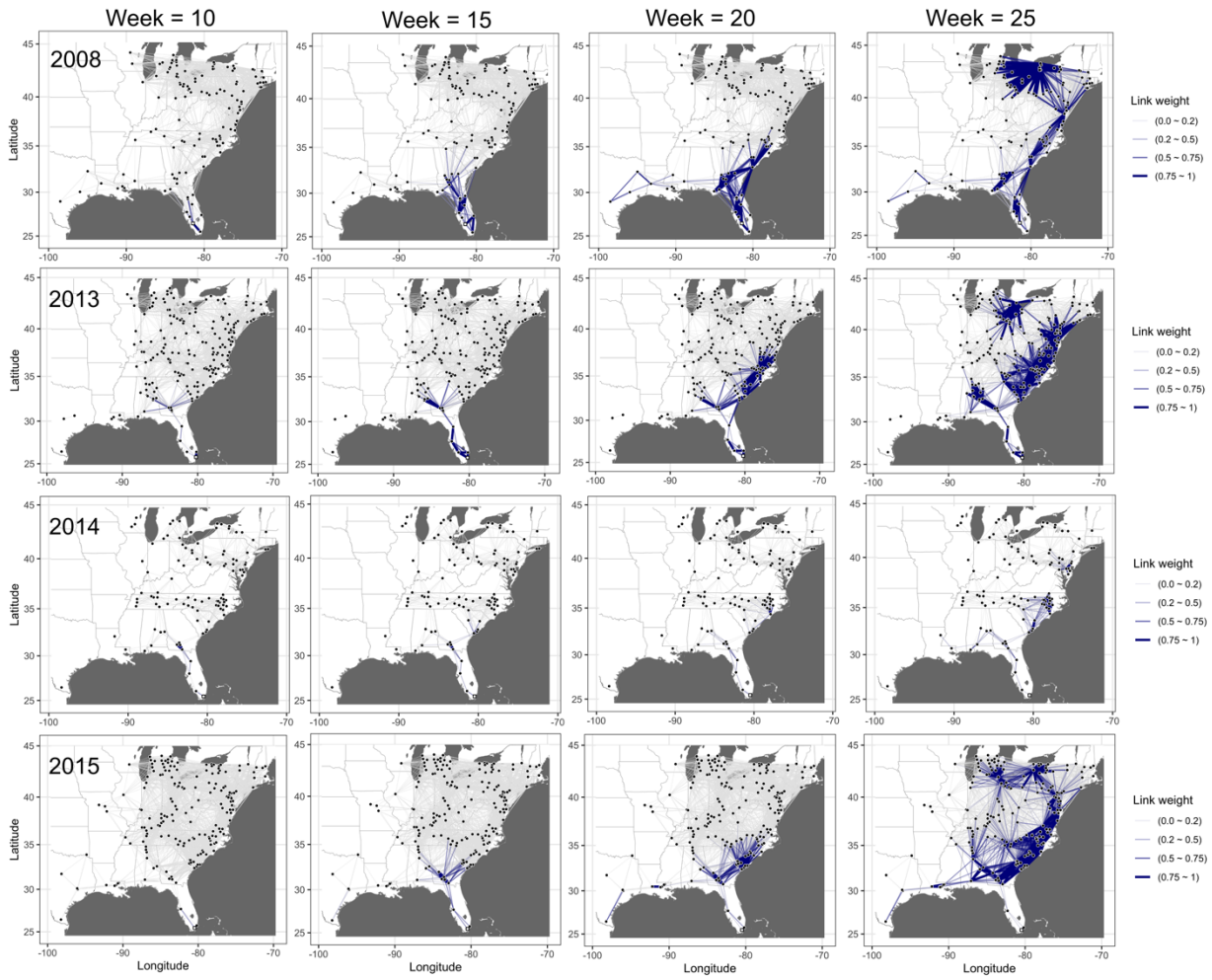


Figure 2.5. Evolving networks resulting from a dynamic network model for cucurbit downy mildew spread in the eastern United States in 2008, 2013, 2014, and 2015. The black circles are the nodes which indicate the locations of disease outbreak. The open square is the initial disease source. The gray links represent the underlying complete static network. The blue links represent the evolving dynamic networks. The dynamic network links differ in width corresponding to the calculated link weight (probabilities), with darker and thicker links indicating greater probabilities of disease spread.

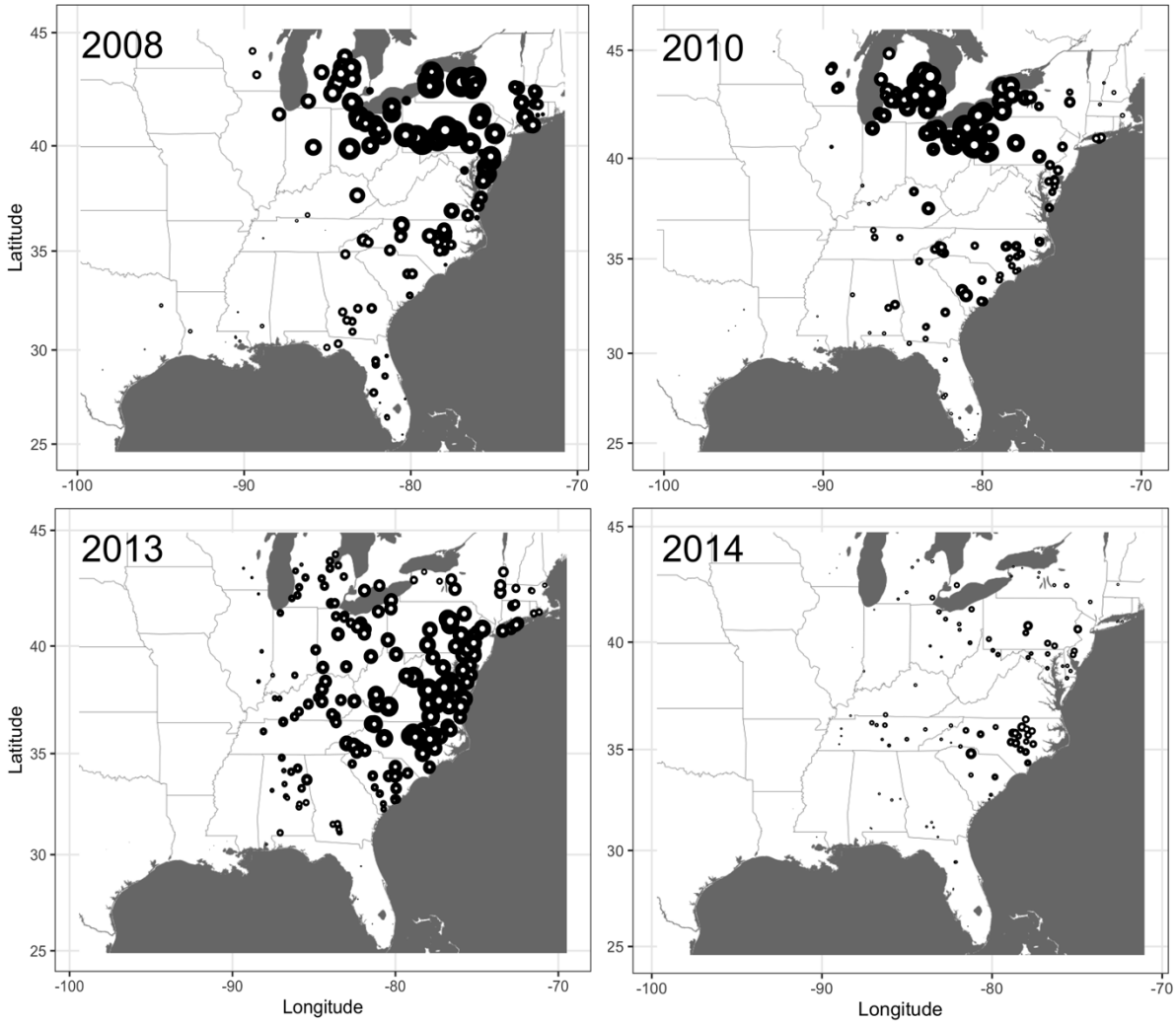


Figure 2.6. The in- and out-degrees calculated for 2008, 2013, 2014, and 2015 networks. The nodes in black are scaled to in-degree (the number of links coming to a node). The nodes in white are scaled to out-degree (the number of links that are coming out from a node defined here as source strength). Nodes represented by large white circles may be strong sources of secondary infection by week 35.

Supplemental Tables

Table S2.1. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2010.

	GC = 1.00		GC = 0.958		GC = 0.958		GC = 0.917		GC = 0.833		GC = 0.548	
Node rank[a]	$\tau = 1.0 \times 10^{-19}$		$\tau = 4.41 \times 10^{-19}$		$\tau = 7.83 \times 10^{-19}$		$\tau = 1.12 \times 10^{-18}$		$\tau = 1.47 \times 10^{-8}$		$\tau = 1.81 \times 10^{-18}$	
1	116 [b]	56 [c]	53	40	61	36	61	30	49	28	44	25
2	**103**	54	61	40	41	33	49	29	44	27	49	25
3	**104**	54	41	39	53	32	44	28	47	26	47	24
4	105	54	48	39	109	30	47	28	57	24	57	24
5	106	54	60	39	49	29	62	28	59	24	59	24
6	**108**	54	109	39	57	29	57	26	61	24	61	23
7	109	54	106	38	59	29	59	26	62	24	38	22
8	**110**	54	111	37	106	29	41	25	38	23	40	21
9	**113**	54	112	37	44	28	45	25	40	22	53	21
10	**114**	54	44	32	47	28	53	25	41	22	62	21
11	61	53	47	32	62	28	55	25	53	22	41	20
12	40	52	49	32	43	27	105	24	106	22	43	20
13	41	52	56	32	55	27	106	24	43	21	55	20
14	48	52	**124**	32	111	27	109	24	45	21	116	20
15	53	52	45	31	112	27	112	24	55	21	45	19
16	60	52	62	31	40	26	38	23	42	20	106	19
17	**107**	52	55	30	45	26	40	23	60	20	109	19
18	112	52	57	30	48	26	60	23	116	20	42	18
19	**122**	52	59	30	56	26	**39**	22	48	19	48	18

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Table S2.2. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2011.

	GC =0.991		GC = 0.991		GC = 0.991		GC = 0.991		GC = 0.931		GC = 0.931	
Node rank[a]	$\tau = 3.79 \times 10^{-13}$		$\tau = 4.1 \times 10^{-13}$		$\tau = 4.41 \times 10^{-13}$		$\tau = 4.72 \times 10^{-13}$		$\tau = 5.03 \times 10^{-13}$		$\tau = 5.34 \times 10^{-13}$	
1	117 [b]	50 [c]	28	48	28	48	109	45	109	45	109	45
2	28	49	109	46	109	46	110	45	110	45	110	45
3	111	49	110	46	110	45	28	43	28	41	111	41
4	109	47	111	45	67	44	111	42	111	41	28	40
5	110	47	44	44	111	44	67	41	44	40	44	38
6	31	46	67	44	44	42	44	40	67	38	31	37
7	29	45	31	43	100	42	31	39	31	37	96	36
8	30	45	48	43	31	41	29	36	45	36	45	35
9	44	45	100	43	29	38	30	36	96	36	71	35
10	48	45	45	42	30	38	45	36	98	36	98	35
11	67	45	27	41	55	38	71	36	5	35	100	35
12	100	45	30	41	59	38	95	36	8	35	5	34
13	27	44	55	41	62	38	96	36	71	35	51	34
14	55	44	59	41	95	38	98	36	95	35	53	34
15	59	44	62	41	98	38	100	36	100	35	66	34
16	62	44	117	41	24	37	5	35	51	34	**72**	34
17	45	43	24	40	25	37	8	35	53	34	**73**	34
18	24	42	25	40	45	37	**9**	35	66	34	95	34
19	117	42	28	40	28	37	109	35	109	34	109	34

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Table S2.3. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2012.

	GC = 0.977		GC = 0.977		GC = 0.936		GC = 0.936		GC = 0.930		GC = 0.711	
Node rank[a]	$\tau = 3.79 \times 10^{-13}$		$\tau = 4.41 \times 10^{-13}$		$\tau = 5.97 \times 10^{-13}$		$\tau = 6.90 \times 10^{-13}$		$\tau = 8.45 \times 10^{-13}$		$\tau = 1.00 \times 10^{-12}$	
1	158 [b]	33 [c]	79	29	139	26	65	25	139	23	64	21
2	81	30	87	29	144	26	62	24	64	22	66	21
3	160	30	90	28	65	25	86	24	65	22	79	21
4	79	29	158	28	81	25	139	24	80	22	80	21
5	82	29	66	27	17	24	17	23	109	22	85	21
6	87	29	81	27	62	24	64	23	66	21	87	21
7	90	29	82	27	66	24	80	23	79	21	92	21
8	16	28	109	27	82	24	98	23	81	21	81	20
9	65	28	139	27	86	24	141	23	82	21	82	20
10	86	28	144	27	92	24	142	23	83	21	83	20
11	109	28	16	26	93	24	144	23	85	21	84	20
12	64	27	64	26	109	24	66	22	87	21	89	20
13	66	27	65	26	141	24	79	22	90	21	90	20
14	85	27	84	26	142	24	82	22	92	21	93	20
15	91	27	86	26	16	23	83	22	93	21	98	20
16	95	27	89	26	64	23	84	22	98	21	139	20
17	**101**	27	95	26	79	23	85	22	141	21	16	19
18	**110**	27	141	26	80	23	89	22	144	21	17	19
19	158	27	79	26	139	23	65	22	139	20	64	19

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Table S2.4. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2013.

	GC = 0.985		GC = 0.946		GC = 0.848		GC = 0833		GC = 0377		GC = 0.377	
Node rank[a]	$\tau = 1.0 \times 10^{-14}$		$\tau = 4.41 \times 10^{-14}$		$\tau = 7.83 \times 10^{-14}$		$\tau = 1.12 \times 10^{-13}$		$\tau = 1.47 \times 10^{-13}$		$\tau = 2.15 \times 10^{-8}$	
1	**203** [b]	70 [c]	**193**	31	194	24	194	21	95	18	83	15
2	**196**	68	194	31	146	22	84	18	100	18	95	15
3	**189**	64	66	30	62	20	88	18	83	17	84	15
4	**191**	64	63	28	63	20	93	18	84	16	100	15
5	**192**	64	90	27	83	20	95	18	88	16	85	14
6	87	62	**94**	27	84	20	96	18	93	16	86	14
7	**98**	62	21	26	88	20	100	18	85	15	88	14
8	**99**	61	62	26	95	20	83	17	86	15	89	14
9	**188**	61	65	26	100	20	146	17	89	15	92	14
10	**190**	61	87	26	86	19	21	16	91	15	93	13
11	**106**	60	95	26	90	19	62	16	92	15	91	13
12	**186**	60	100	26	96	19	63	16	96	15	96	13
13	**198**	60	**145**	26	144	19	85	16	110	15	110	13
14	63	59	84	25	**147**	19	86	16	149	15	149	13
15	65	59	88	25	**19**	18	89	16	21	14	21	13
16	85	59	91	25	21	18	90	16	61	14	61	12
17	86	59	**108**	25	65	18	91	16	62	14	62	12
18	89	59	144	25	66	18	92	16	90	14	87	12
19	**203**	59	**193**	25	194	18	194	16	95	14	83	12

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Table S2.5. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2014.

	GC = 0.991		GC = 0.991		GC = 0.991		GC = 0.991		GC = 0.807		GC = 0.807	
Node rank[a]	$\tau = 2.55 \times 10^{-17}$		$\tau = 3.17 \times 10^{-17}$		$\tau = 3.48 \times 10^{-17}$		$\tau = 3.79 \times 10^{-17}$		$\tau = 5.34 \times 10^{-17}$		$\tau = 5.97 \times 10^{-8}$	
1	46 [b]	26 [c]	46	24	46	23	46	23	57	20	51	19
2	90	26	57	24	57	23	57	23	51	19	57	19
3	95	26	51	22	51	22	51	21	56	18	56	18
4	51	25	83	22	95	22	90	21	58	18	58	18
5	56	25	90	22	90	21	39	20	40	17	40	17
6	57	25	95	22	39	20	44	20	41	17	41	17
7	58	24	39	21	42	20	48	20	42	17	42	17
8	83	24	48	21	44	20	55	20	44	17	44	17
9	87	24	42	20	48	20	56	20	46	17	48	17
10	39	23	44	20	55	20	95	20	48	17	55	17
11	40	22	55	20	56	20	42	19	55	17	31	16
12	44	21	56	20	58	19	58	19	31	16	39	16
13	48	21	61	20	61	19	61	19	39	16	43	16
14	55	21	58	19	83	19	83	19	43	16	45	16
15	61	21	87	19	40	18	40	18	45	16	46	16
16	41	20	40	18	54	18	54	18	52	16	52	16
17	42	20	43	18	**60**	18	31	17	53	16	53	16
18	84	20	52	18	31	17	41	17	54	16	54	16
19	52	19	53	18	41	17	43	17	59	16	59	16

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Table S2.6. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2015.

	GC = 1.00		GC = 0.913		GC = 0.909		GC = 0.895		GC = 0.799		GC = 0.249	
Node rank[a]	$\tau = 1.0 \times 10^{-12}$		$\tau = 4.41 \times 10^{-12}$		$\tau = 7.83 \times 10^{-12}$		$\tau = 1.12 \times 10^{-11}$		$\tau = 1.47 \times 10^{-11}$		$\tau = 1.81 \times 10^{-11}$	
1	**207** [b]	90 [c]	**172**	39	13	27	13	24	13	24	13	19
2	**208**	87	**170**	38	152	27	110	24	12	19	12	18
3	**99**	82	**173**	37	155	27	160	23	60	19	110	17
4	**182**	81	**162**	33	157	27	12	22	71	18	148	17
5	**175**	80	**180**	33	160	27	155	22	110	18	174	17
6	**206**	79	55	31	148	26	71	21	183	18	183	17
7	**93**	78	**73**	31	12	25	152	21	146	17	186	17
8	**176**	78	**156**	31	110	25	148	20	148	17	**187**	17
9	55	77	67	30	146	24	157	20	152	17	**188**	17
10	**149**	77	146	30	55	23	**53**	19	155	17	60	16
11	**94**	76	148	30	60	23	55	19	167	17	146	16
12	**96**	73	**151**	30	**66**	22	60	19	168	17	155	16
13	**163**	73	152	30	**70**	22	69	19	**169**	17	**166**	16
14	**89**	72	**154**	30	71	22	164	19	174	17	167	16
15	**158**	72	155	30	**62**	21	168	19	177	17	**171**	16
16	**159**	72	164	30	67	21	177	19	**181**	17	**178**	16
17	**100**	71	13	29	69	21	**10**	18	184	17	184	16
18	**134**	71	**57**	29	**72**	21	**11**	18	186	17	**185**	16
19	**207**	71	**172**	29	13	21	13	18	13	17	13	15

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Table S2.7. Ranking of nineteen most important nodes based on degree centrality for networks generated using a combination of values of the proportion of nodes in the giant component (GC) and connectivity threshold (τ) for epidemic data in 2016.

Node rank[a]	GC = 0.967 $\tau = 1.0 \times 10^{-12}$		GC = 0.951 $\tau = 1.47 \times 10^{-12}$		GC = 0.951 $\tau = 1.95 \times 10^{-12}$		GC = 0.951 $\tau = 2.42 \times 10^{-12}$		GC = 0.951 $\tau = 2.89 \times 10^{-12}$		GC = 0.951 $\tau = 3.37 \times 10^{-12}$	
1	**154** [b]	48 [c]	211 [b]	39	79	32	93	30	93	29	91	29
2	**55**	47	201	39	74	32	82	29	109	29	89	29
3	**54**	46	**129**	38	201	32	88	29	92	27	75	27
4	**53**	46	128	38	159	31	91	28	91	26	109	24
5	**50**	46	**81**	38	109	31	89	28	89	26	83	24
6	**48**	46	**133**	37	93	31	83	28	83	24	76	23
7	**164**	45	**126**	37	82	30	76	28	82	23	95	23
8	**193**	45	**124**	35	212	30	75	28	76	23	92	23
9	213	44	**87**	35	136	30	136	28	75	23	90	23
10	**206**	44	212	35	118	29	92	28	90	23	88	21
11	152	44	208	33	114	29	90	26	88	22	93	21
12	**166**	43	**204**	32	90	29	79	26	125	22	82	21
13	**155**	43	136	31	76	29	127	24	**97**	22	80	20
14	163	43	**135**	31	208	29	109	24	95	21	79	20
15	**132**	42	118	31	130	28	95	24	79	21	128	20
16	**165**	42	213	31	127	28	158	23	74	20	125	20
17	156	41	160	31	**111**	28	130	23	128	20	119	18
18	160	40	159	31	92	27	125	22	121	20	117	18
19	157	40	158	31	91	27	121	22	119	20	114	18

[a] Numerical rank of the nineteen most important nodes in the network based on degree centrality.
[b] Value refers to the actual number assigned to a node within the network; Values in regular font are for nodes that were ranked as important by all the connectivity thresholds tested, while those in bold font are for nodes that are unique and were not ranked as important by all the thresholds tested.
[c] The calculated degree centrality value.

Supplemental Figures

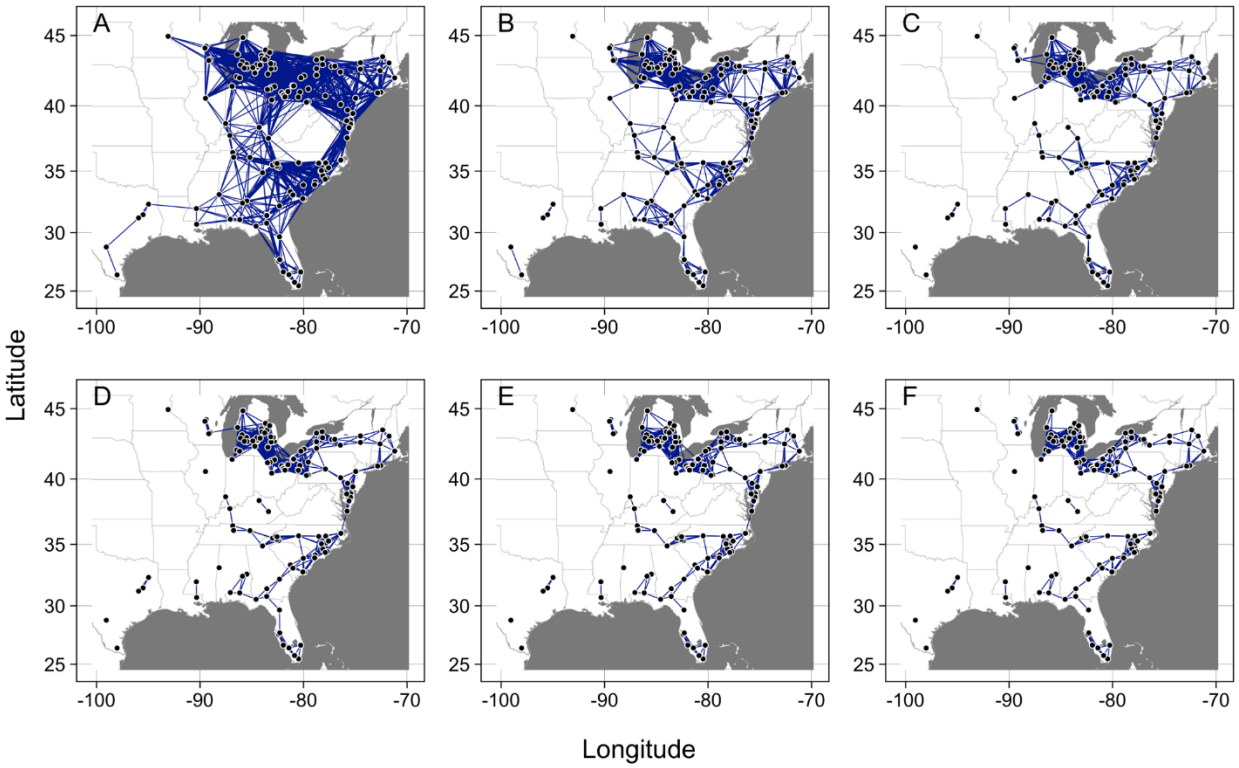


Figure S2.1. Examples of 2010 networks generated with spread parameter b = 3.36. The thresholds for the networks shown in panels A - F are 1×10^{-19}, 4.41×10^{-19}, 7.83×10^{-19}, 1.12×10^{-18}, 1.47×10^{-18}, and 1.81×10^{-18}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

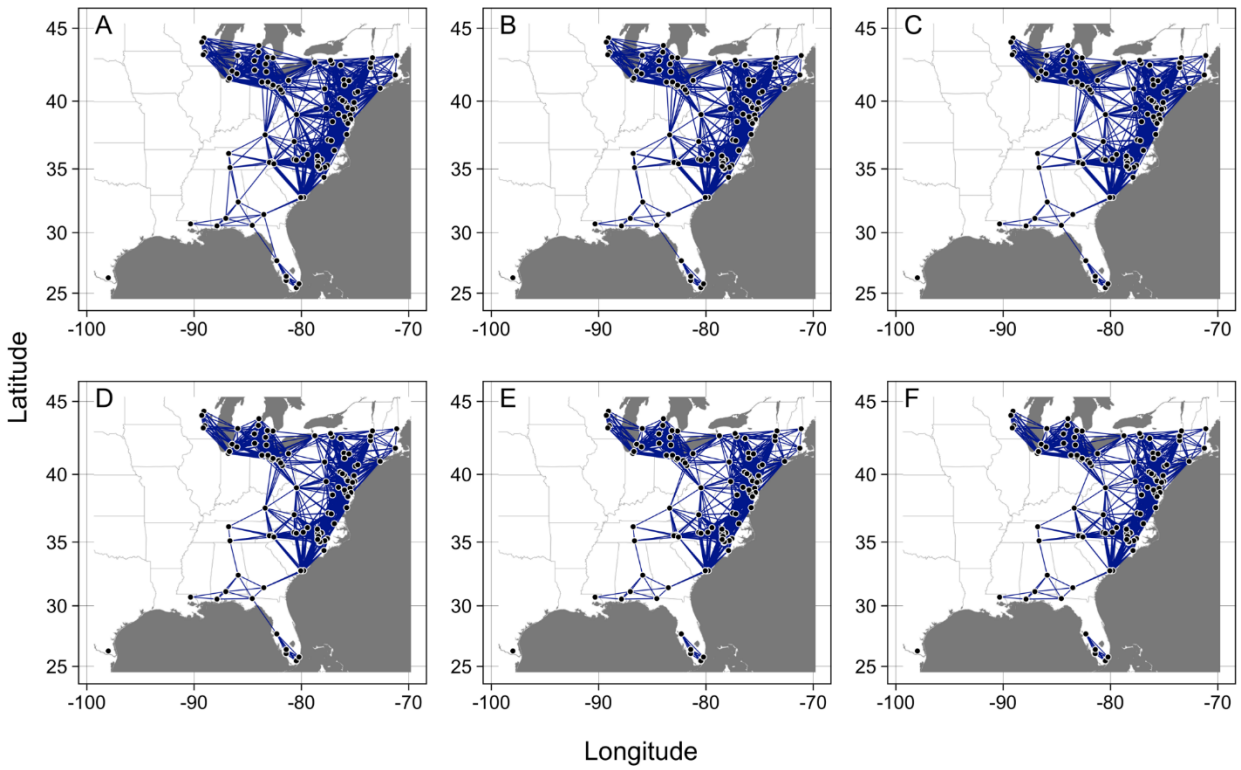


Figure S2.2. Examples of 2011 networks generated with spread parameter *b* = 2.2. The thresholds for the networks shown in panels A - F are 3.79×10^{-13}, 4.1×10^{-13}, 4.41×10^{-13}, 4.72×10^{-13}, 5.03×10^{-13}, and 5.34×10^{-13}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

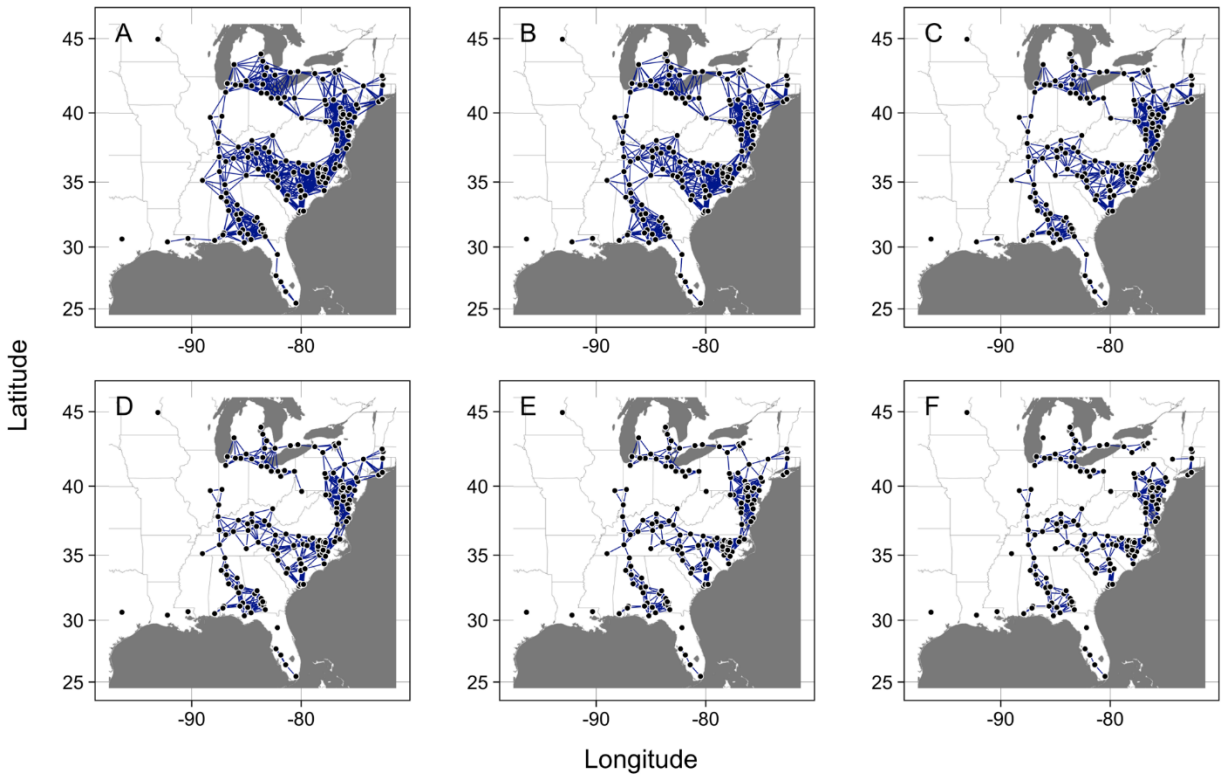


Figure S2.3. Examples of 2012 networks generated with spread parameter *b* = 2.32. The thresholds for the networks shown in panels A - F are 3.79×10^{-13}, 4.41×10^{-13}, 5.97×10^{-13}, 6.9×10^{-13}, 8.45×10^{-13}, and 1×10^{-12}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

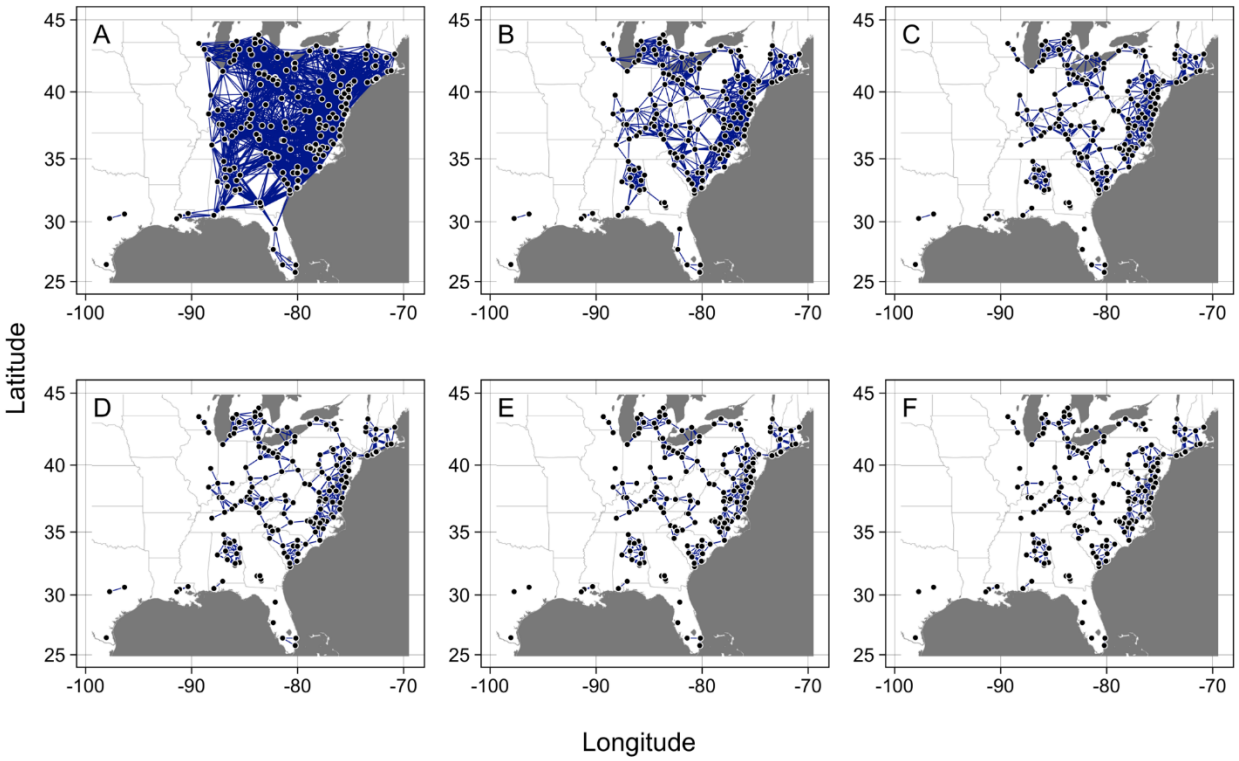


Figure S2.4. Examples of 2013 networks generated with spread parameter *b* = 2.51. The thresholds for the networks shown in panels A - F are 1×10^{-14}, 4.41×10^{-14}, 7.83×10^{-14}, 1.12×10^{-13}, 1.47×10^{-13}, and 1.81×10^{-13}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

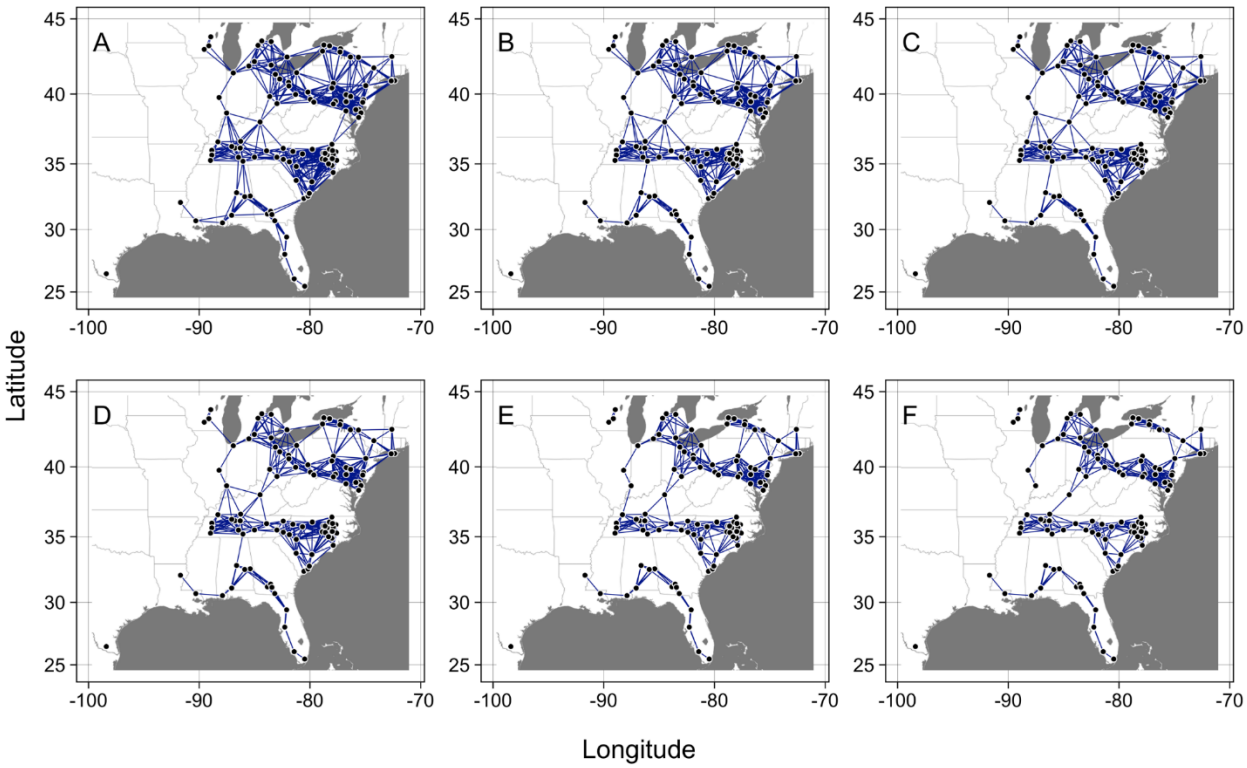


Figure S2.5. Examples of 2014 networks generated with spread parameter b = 3.02. The thresholds for the networks shown in panels A - F are 2.55×10^{-17}, 2.86×10^{-17}, 3.17×10^{-17}, 3.48×10^{-17}, 3.79×10^{-17}, and 5.34×10^{-17}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

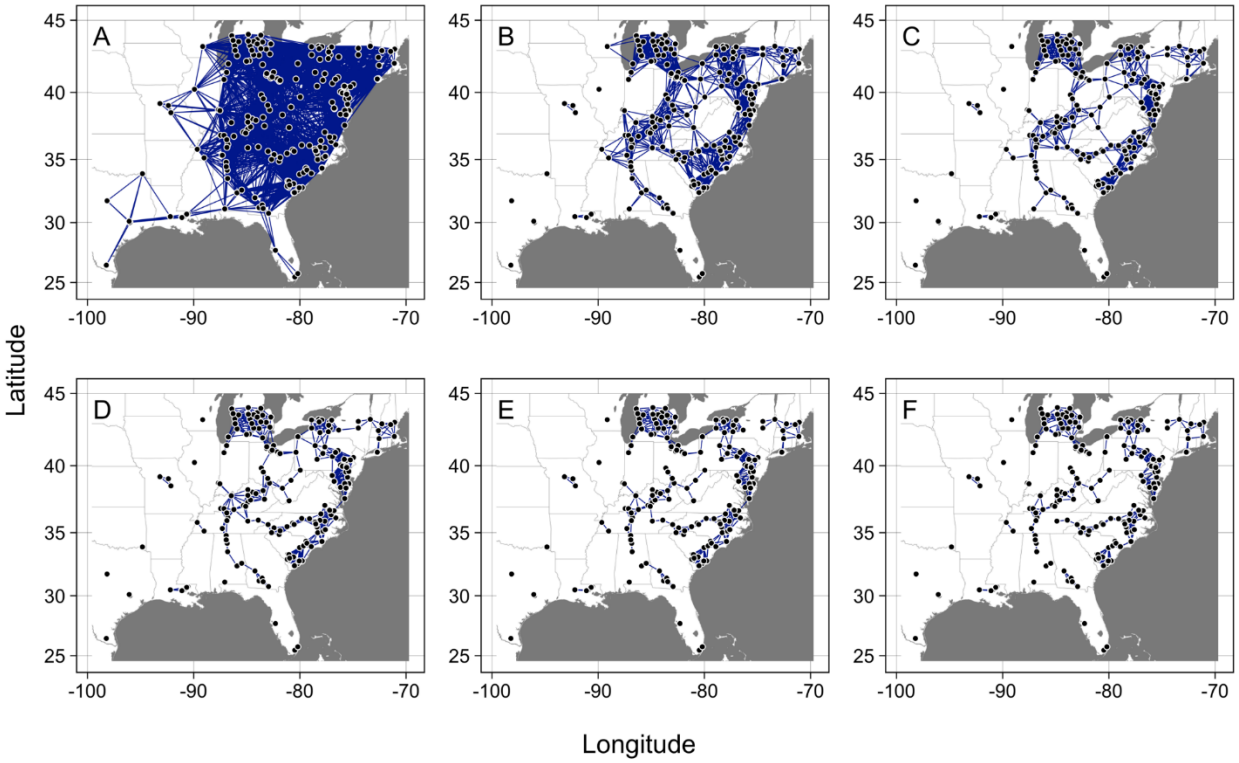


Figure S2.6. Examples of 2015 networks generated with spread parameter *b* = 2.11. The thresholds for the networks shown in panels A - F are e 1×10^{-12}, 4.41×10^{-12}, 7.83×10^{-12}, 1.12×10^{-11}, 1.47×10^{-11}, and 1.81×10^{-11}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

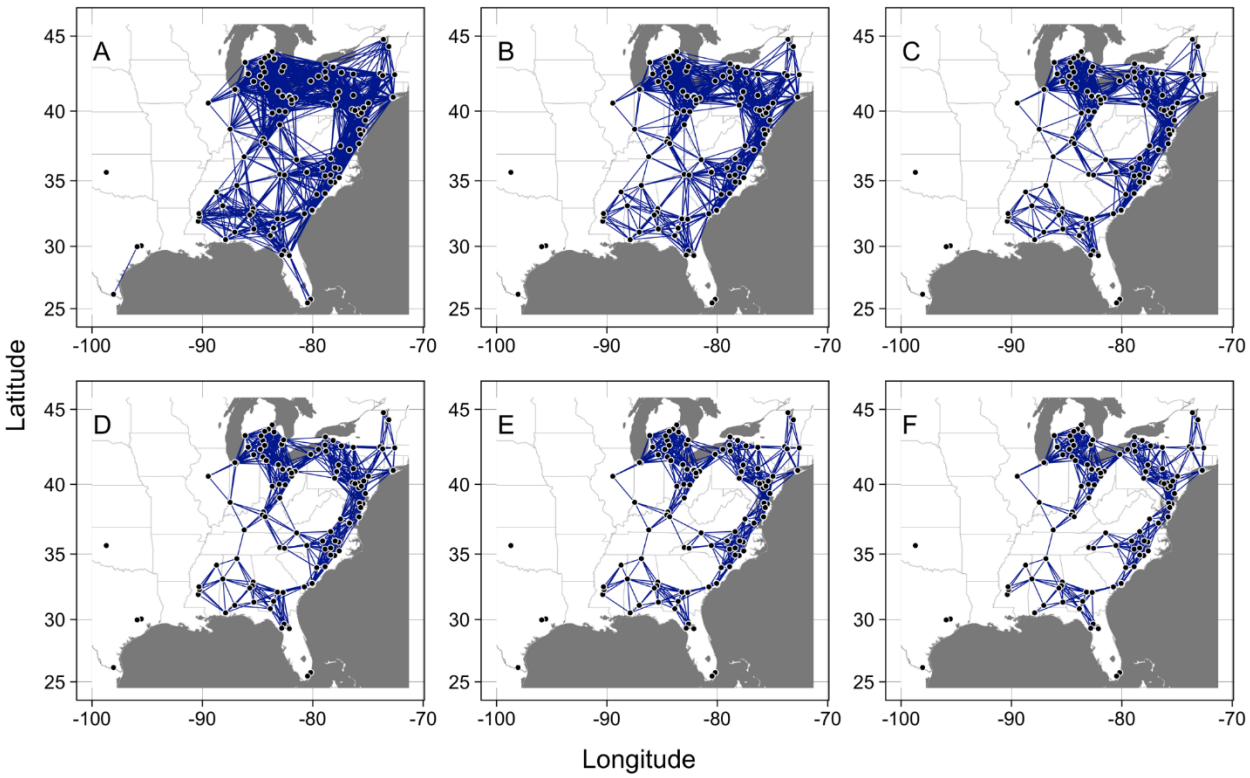


Figure S2.7. Examples of 2016 networks generated with spread parameter *b* = 2.11. The thresholds for the networks shown in panels A - F are 1×10^{-12}, 1.47×10^{-12}, 1.95×10^{-12}, 2.42×10^{-12}, 2.89×10^{-12}, and 3.37×10^{-12}, respectively. When the threshold is minimal, everything is connected. As the threshold increases, links reduce between nodes that are far apart. When the threshold is high, links only occur between nodes that are close to each other.

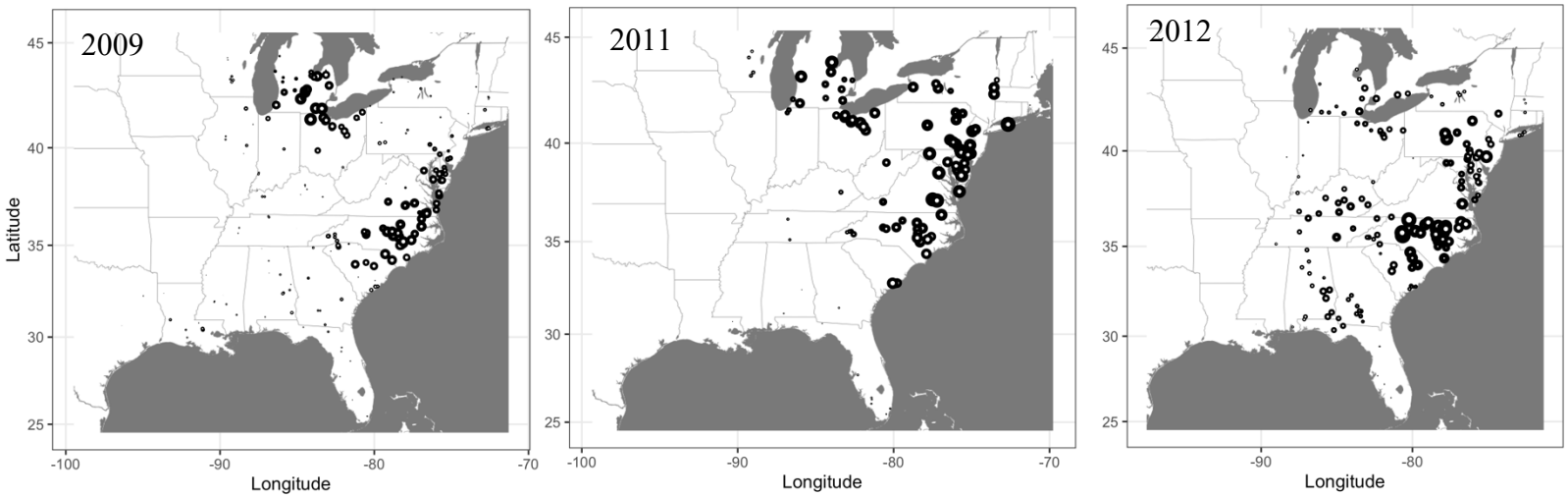


Figure S2.8. The in- and out-degrees calculated for 2009, 2011, and 2012 networks. The nodes in black are scaled to in-degree (the number of links coming to a node). The nodes in white are scaled to out-degree (the number of links that are coming out from a node defined here as source strength). Nodes represented by large white circles may be strong sources of secondary infection by week 35.

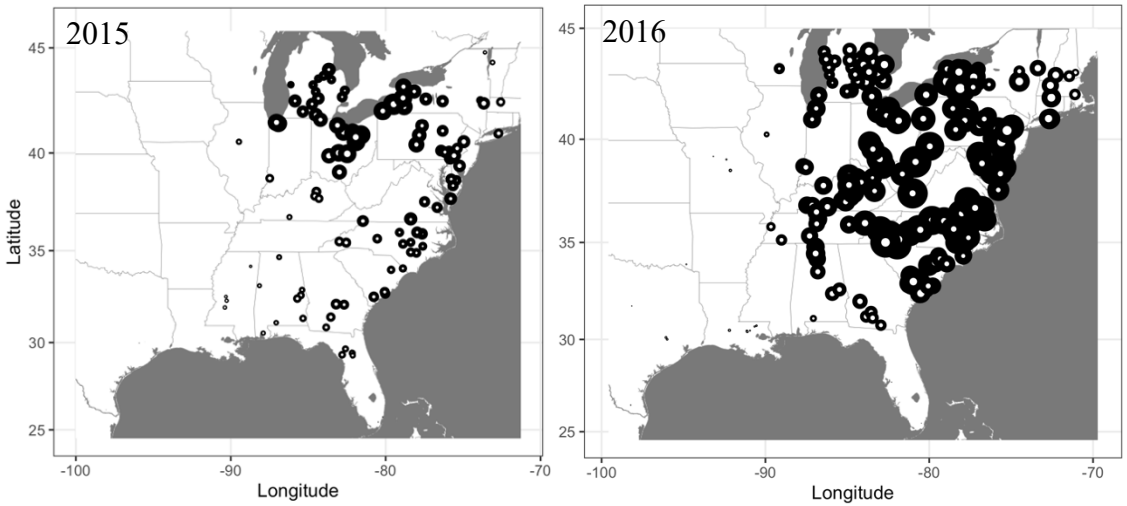


Figure S2.9. The in- and out-degrees calculated for 2015 and 2016 networks. The nodes in black are scaled to in-degree (the number of links coming to a node). The nodes in white are scaled to out-degree (the number of links that are coming out from a node defined here as source strength). Nodes represented by large white circles may be strong sources of secondary infection by week 35.

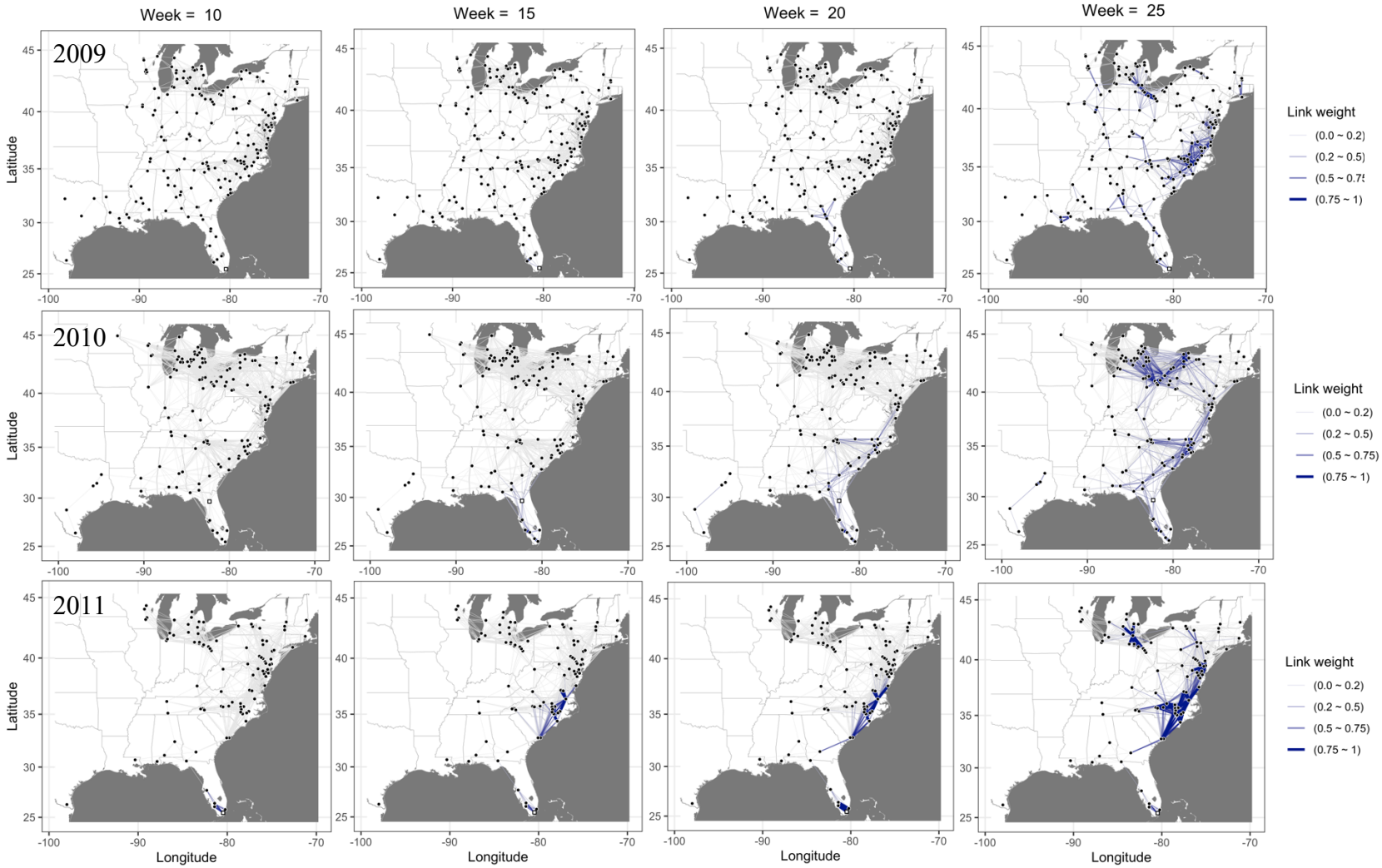


Figure S2.10. Evolving networks resulting from a dynamic network model for cucurbit downy mildew spread in the eastern United States in 2009, 2010, and 2011. The black circles are the nodes which indicate the locations of disease outbreak. The open square is the initial source of disease. The gray links represent the underlying complete static network. The blue links represent the evolving dynamic networks. The dynamic network links differ in width corresponding to the calculated link weight (probabilities), with darker and thicker links indicating greater probabilities of disease spread.

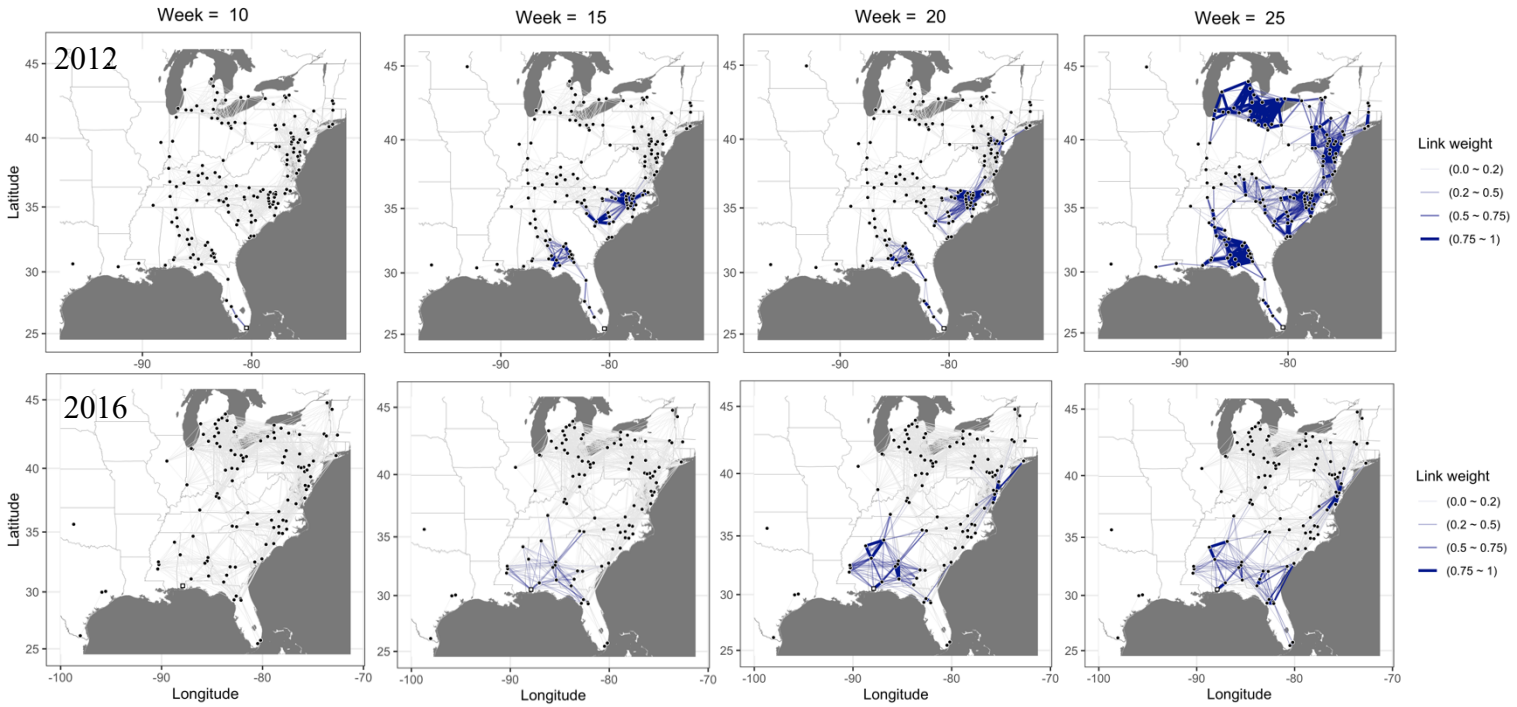


Figure S2.11. Evolving networks resulting from a dynamic network model for cucurbit downy mildew spread in the eastern United States in 2012 and 2016. The black circles are the nodes which indicate the locations of disease outbreak. The open square is the initial source of disease. The gray links represent the underlying complete static network. The blue links represent the evolving dynamic networks. The dynamic network links differ in width corresponding to the calculated link weight (probabilities), with darker and thicker links indicating greater probabilities of disease spread.

CHAPTER 3

Network Analysis of the Spread of Cucurbit Downy Mildew in the Eastern United States: Identifying Highly Connected Locations for Risk-Based Surveillance and Disease Control

Abstract

Surveillance is critical in the rapid implementation of control measures for diseases caused by aerially dispersed plant pathogens, but such programs can be resource-intensive, especially for long-distance dispersed pathogens. The current platform for monitoring, predicting, and communicating the risk of cucurbit downy mildew is expensive to maintain. In this study, we focused on finding fields for surveillance and treatment because knowing where to monitor for the disease could reduce surveillance costs and knowing where to treat could slow down the invasion process during the growing season. We constructed static and dynamic networks using epidemic data collected from 2008 to 2016 and used three strategies to identify these fields. First, we modeled the probabilities of different nodes being infected over discrete weekly time steps within a year. Secondly, we selectively removed nodes from a network and calculated probabilities. Finally, we analyzed recurring patterns (infection frequency) across the years. Betweenness centrality (BWC) was the most useful measure in identifying the most important fields compared to other centrality measures examined. Also, degree centrality, a commonly used measure of more central nodes, was not as effective as BWC or other centrality measures assessed. Fields in Maryland, North Carolina, Ohio, South Carolina, and Virginia were the most central in the network. Further, removing fields identified as important based on BWC limited the risk of disease spread based on a dynamic network model incorporating a power-law function for pathogen dispersal. Combining the dynamic network model and centrality measures helped identify the highly connected cucurbit fields in the southeastern United States and the mid-Atlantic region. These highly connected fields may be used to inform surveillance and strategies for controlling cucurbit downy mildew in the eastern United States.

Keywords: cucurbit downy mildew, centrality measures, infection frequency, probability

1. Introduction

Pathogen dispersal is a fundamental property in developing disease epidemics at different spatial scales that range from the local to landscape level. The transmission of invasive plant pathogens and the spread of resultant epidemics influences essential ecosystem services, including biodiversity and food production in agricultural systems (Brown and Hovmøller, 2002; Crowl et al., 2008). Measures that involve containment and eradication programs can be implemented to reduce the potential impact of these epidemics. However, the planning and implementation of any specific measure will require an understanding of the mechanics of invasions and the ecological consequences, risks, and dynamics of disease spread. Such efforts can benefit greatly from epidemic records within a region as they enable an analysis of the overall structure of pathogen dispersal. Information from such analyses can help design control programs for disease epidemics and risk-based surveillance. For example, the timely recording of animal movements was fundamental in the containment of the 2011 foot and mouth disease epidemic in the UK, for which retrospective analyses demonstrated that initial spread was influenced by the frequency of animal movement (Ferguson et al., 2001; Kao et al., 2006).

One approach to understand pathogen dispersal and spread of resultant disease epidemics is through network analysis, a method that is becoming increasingly popular but of limited application in plant disease epidemiology (Garrett et al., 2018; Xing et al., 2020). Networks consist of ‘nodes’ and ‘links’, where nodes are the entities of interest (e.g., individual fields or observed locations of disease outbreak), while links serve to connect the nodes in various ways, for example, the potential of encounter or disease transmission between two nodes. Further, networks can be weighted with link weights assumed to be proportional to the probability of transmission. Networks have been constructed to describe pathogen dispersal and spread of epidemics caused

by aerially dispersed plant pathogens such as *Phakopsora pachyrhizi* in soybean (Sutrave et al., 2012; Sanatkar et al., 2015) and *Podosphaera macularis* in the hop (Gent et al., 2019). The primary determinants of a pathogen's dispersal such as source strength, location of host populations, and relevant covariate information can be formulated as a network spreading model (Gent et al., 2019; Firester et al., 2018; Garrett et al., 2018; Sanatkar et al., 2015; Sutrave et al., 2012). Such models combine the spatial and dynamic components of an epidemic and provide the underlying contact structure of landscape connectivity (With et al., 1997).

The choice of networks to be studied depends on several factors that include the disease of interest and specific questions. The latter will, in turn, influence the type of network measures to be used in the analysis of pathogen dispersal and disease spread. Static and dynamic networks are common in landscape connectivity analyses. Static networks have structures that do not change, whereas dynamic networks have structures that change over time. Both static and dynamic networks are applicable in plant disease epidemiology (e.g., Sanatkar et al., 2015; Sutrave et al., 2012). In dynamic networks, between-node distances, host availability, wind speed, and wind direction can formulate a susceptible-infected (SI) model to describe epidemic spread (Sutrave et al., 2012).

Further, plant diseases display seasonal differences in the occurrence and intensity of epidemics (Campbell and Madden, 1990). Thus, analysis of data from an entire year and even over multiple years will be necessary to determine if there are recurring patterns that could be useful for designing effective disease control measures. There are several different measures in the network analysis of pathogen dispersal and disease spread. Some of these measures provide general descriptions of the network structure (Keeling and Eames, 2005), while other measures can apply to surveillance and management of epidemics (Ferrari et al., 2014). For example,

centrality measures such as betweenness, closeness, degree, and eigenvector help assess the number of contacts and identify specific nodes in the highly connected network (Gent et al., 2019; Sutrave et al., 2012). Highly connected nodes provide more effective surveillance and opportunities for more targeted control to reduce disease spread within the network. An open question still exists as to which centrality measures are most important for identifying important nodes for surveillance and managing real-world networks (Holme, 2017). Due to inherent differences in pathogen dispersal and disease spread mechanisms, centrality measures used to identify essential nodes for surveillance are specific to different pathosystems (Holme, 2018).

A motivating plant disease example for network analysis to inform surveillance and disease control is cucurbit downy mildew (CDM), a pathosystem of interest for several reasons. A resurgence of the disease occurred around the world in the last 20 years, which fundamentally influenced cucurbit production and disease management at multiple scales (Ojiambo et al., 2015). The disease is caused by an obligate pathogen, *Pseudoperonospora cubensis*, that exhibits significant long-distance dispersal (Ojiambo et al., 2011). The resurgence of CDM in Europe and the United States has been attributed to the introduction of a new pathotype of the pathogen that was previously limited to East Asia (Cohen et al., 2015; Thomas et al., 2017).

Fungicides are integral to CDM control due to the lack of cultivars with adequate resistance, and in the absence of control measures, the disease can result in complete crop loss (Holmes et al., 2015). The pathogen overwinters below the 30-degree latitude in southern Florida in the continental United States, and disease occurrence in northern states relies on pathogen dispersal from the south. In 2008, surveillance of disease occurrence based on a series of sentinel and non-sentinel plots was implemented as part of the CDM ipmPIPE (https://www.ipmpipe.org/) (Ojiambo et al., 2011). Based on the prediction framework developed by Main et al. (2001) and

the sentinel plot data, an integrated aerobiological model was developed to predict the disease occurrence and progression in the eastern United States (Neufeld et al., 2018) to guide growers on when to apply the initial spray. Surveys conducted in Georgia, Michigan, and North Carolina show that the forecasting system resulted in an average reduction of two to three fungicide applications than calendar-based application schedules (Ojiambo et al., 2011). This reduction in fungicide applications translates to > $6 million savings to the cucurbit producers in these three states alone annually. Due to limitations in resources for disease surveillance (Ojiambo et al., 2011), there is increasing interest to identify locations that are critical for pathogen dispersal and disease spread within the region.

This study builds upon recent work conducted to characterize the network structure of the dispersal of *P. cubensis* and the spread of cucurbit downy mildew in the eastern United States (Chapter 2). We specifically focus on identifying the centrality measures that can be directly applicable for surveillance and management of cucurbit downy mildew to identify highly connected nodes. When combined with frequency-based selection, these centrality measures can be used to reduce the resources required to survey and predict epidemic progress (Sutrave et al., 2012). A dynamic network model for cucurbit downy mildew in the eastern United States is developed using epidemic data, wind speed, and wind direction based on a modeling framework described by Sutrave et al. (2012). The latter is essentially an SI model that describes disease spread between nodes in the eastern United States wherein nodes are observed locations and links are possible transmission routes from infected nodes to susceptible nodes. The specific objectives of this study were to

1. Identify centrality measures that are useful for surveillance and control of cucurbit downy mildew.

2. Identify highly connected nodes that are critical in the pathogen dispersal and spread of cucurbit downy mildew.
3. Determine how the removal of highly connected nodes influences the spread and containment of cucurbit downy mildew in the eastern United States.

Due to differences in patterns of pathogen dispersal and disease spread between years, data are analyzed from multiple years to generate robust findings and identify any recurring patterns that could be useful in disease control.

2. Methods

2.1. Data source

Cucurbit downy mildew epidemic records in the eastern United States from 2008 to 2016 were used in this study. The data was obtained from the CDM ipmPIPE database (http://cdm.ipmpipe.org) that tracks reports of CDM occurrences in the United States (Ojiambo et al., 2011). Epidemic records in the system include reports from a network of regularly monitored plots (sentinel plots) and voluntary reports (non-sentinel plots) submitted by commercial growers, agricultural researchers, and the general public. Sentinel plots were strategically placed within specific states and planted with different cucurbit host types for monitoring CDM occurrences. During this study period (2008 - 2016), the sentinel plots were located at research facilities or commercial fields with standard dimensions of 15 m × 61 m and were georeferenced using the Global Positioning System. These plots were planted early and regularly monitored for disease symptoms every 1-2 weeks by state collaborators and extension specialists. The cucurbits grown in the sentinel plots were cucumber cv. Straight 8 and Poinsett 76 (*Cucumis sativus*), cantaloupe cv. Hales Best Jumbo (*Cucumis melo*), acorn squash cv. Table Ace (*Cucurbita pepo*), butternut

squash cv. Waltham (*Cucurbita moschata*), giant pumpkin cv. Big Max (*Cucurbita maxima*), and watermelon cv. Micky Lee (*Citrullus lanatus*) (Ojiambo et al., 2011).

Voluntary reports were from locations not designated for regular surveillance, i.e., commercial fields, research plots, and home gardens (Table 3.1). The appearance of infection triggered these voluntary reports (non-sentinel plot reports). The reports are essential for many reasons. First, CDM was reported earlier in the non-sentinel plots in some years (e.g., 2011, 2013, 2015, and 2016) before the sentinel plots (Figure 3.1). These early reports are critical for inferring source attribution and CDM spread. Secondly, this data is available. It is expensive to establish and monitor sentinel plots in all states in the eastern United States (Ojiambo et al., 2011). Due to this resource limitation and sentinel data sparsity, including these non-sentinel reports allows for much data as possible for analysis. Third, since *P. cubensis* is an aerially dispersed pathogen, it is crucial to include all locations with CDM to infer the epidemic extent, identify possible disease hops, and capture information from locations with no sentinel plots.

The latitudes and longitudes for the sentinel and non-sentinel plots were generated from the customized section of the CDM ipmPIPE website. The latitudes and longitudes of county centroids - extracted from US Census Bureau 1990 Gazetteer Files - were used as the approximate georeferenced points where no plot data were available. The compiled data from sentinel and non-sentinel plots included the date of first disease symptoms, month, affected host type, planting type, disease incidence, state, county, and location. The total number of disease cases across the study period ranged from 114 to 220, while the number of counties affected ranged from 86 to 179 (Table 3.1). A correlation was done to determine how these numbers were impacted by the number of plots and counties with active surveillance in each year's data set (Figure S2.2). A disease case

represented a unique combination of host and date of first disease symptoms at a particular location.

Hourly wind speed and direction at each sentinel plot were derived from weather observations in the national oceanic and atmospheric administration integrated surface database (Smith et al., 2011) provided by BASF (Research Triangle Park, Raleigh, NC). The wind measurements were taken at the height of 10 m. The wind direction is the direction the wind is blowing from, e.g., wind coming from the north is a northerly wind, and a southerly wind is a wind coming from the south. The raw observations for the meteorological wind direction for a north wind is 360°, a south wind is 180°, a west wind is 270°, and an east wind is 90° (Figure S3.1A). The wind direction (*wd*) in degrees was converted to a mathematical direction (*md*) in degrees using the formula:

$$md = \begin{cases} 270 - wd \\ 360 + (270 - wd) \end{cases} \quad (1)$$

This mathematical convention for the meteorological wind direction implies that a north wind is 270°, a south wind is 90°, a west wind is 0°, and an east wind is 180° (Figure S3.1B). The mathematical direction vector points in the direction that wind will transport the particles, i.e., a meteorological north wind (360°) will blow a particle to the south (270° mathematical). The mathematical direction in degrees was converted to radians. The x and y (u and v) components of the hourly wind vectors were then calculated as $x = rcos\Theta$ and $y = rsin\Theta$ where *r* is the wind speed in miles per hour and Θ is the wind direction in radians (Figure S3.1C).

2.2. Static network analysis

Spatial networks were constructed for each epidemic year to provide insights into the structure of CDM spread in the eastern United States. The general methodology involved

positioning node *i* at a specific location, then connecting node *i* and node *j* using a probability based on the distance between the two nodes. The probability is given by a connection kernel which usually decays with distance such that connections are predominantly localized (Danon et al., 2010). In this study, nodes were a combination of sentinel plots, research fields, home gardens, and commercial fields in the eastern United States. Other locations in the eastern United States that were not monitored in this study may contribute to the risk and spread of CDM. However, only the reported locations to have CDM were included in this analysis since this is the available data, one reason why we combined both sentinel and non-sentinel plots.

In this study, a link between two nodes was determined as a function of distance. The between-node Euclidean distances were calculated using the Haversine formula (Sinnot, 1984) and implemented in the R programming language (R Development Core Team) using the package geosphere (Hijmans, 2017). The *x* and *y* displacement vectors for two nodes were calculated based on equirectangular projection. The formulas are:

$$\begin{aligned} \mathrm{z} &= \sin^2\left(\frac{\varphi_\mathrm{j}-\varphi_\mathrm{i}}{2}\right) + \cos(\varphi_\mathrm{j})\cos(\varphi_\mathrm{i})\sin^2\left(\frac{\lambda_\mathrm{j}-\lambda_\mathrm{i}}{2}\right) \\ l_{ij} &= \mathrm{R}\times 2\times \mathrm{atan2}\left(\sqrt{\mathrm{z}},\sqrt{1-\mathrm{z}}\right) \\ x &= \mathrm{R}\times(\lambda_\mathrm{j}-\lambda_\mathrm{i})\cos\left(\frac{\varphi_\mathrm{j}+\varphi_\mathrm{i}}{2}\right) \\ y &= \mathrm{R}\times(\varphi_\mathrm{j}-\varphi_\mathrm{i}) \end{aligned} \qquad (2)$$

where φ = latitude (radians), λ = longitude (radians), R = radius of the earth (mean = 6,371,000 m), and l_{ij} = haversine distance between node *i* to node *j*.

The links were created using an inverse power-law dispersal kernel $y = \left(l_{ij}\right)^{-b}$, where y is the probability of transmission from node *i* to node *j* (Andersen et al., 2019), l_{ij} is the distance between node *i* to node *j*, while *b* is the spread parameter (Ojiambo et al., 2017). Values for parameter *b* used in this study ranged from 1.51 to 3.36 and these were obtained from a previous study on the isotropic spread of CDM in the eastern United States from 2008 to 2016 (Ojiambo et

al., 2017). A link was created between node *i* to node *j* if y > τ for some arbitrary threshold values τ. A simpler way to think about $y = \left(l_{ij}\right)^{-b}$ and τ is that we connect node *i* to node *j* based on whether they are within a certain distance of each other.

Two analyses of static networks were done. First, several static networks were created for different τ values (sections 2.2.1 and 2.2.2). Then a single network where each node was connected to at least another node was chosen for individual analysis in sections 2.2.1 and 2.3.

2.2.1. Centrality measures

Four centrality measures, namely, betweenness, closeness, degree, and eigenvector, were assessed to determine their usefulness in surveillance and management of CDM in the eastern United States (Meghanathan and Lawrence, 2016). These centrality measures have been used in network analysis of aerially dispersed plant pathogens and have relevance in epidemic spread (Table 3.2). Betweenness centrality (BWC) quantifies the number of shortest paths from each node *i*, to every other node *j*, that run through a focal node *k*: $BWC_k = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$, where g_{ij} is the number of paths from node *i* to node *j* and g_{ikj} is the number of paths that run through node *k*. Nodes with high BWC scores are used more often than nodes with lower scores and thus, more important to spread across the network. Closeness centrality (CLC) measures how close (on average, in number shortest paths) a node is to other nodes, and thus how quickly an epidemic starting at that node might infect a large proportion of the network: $CLC_i = \frac{1}{\sum_j l_{ij}}$, where l_{ij} is the shortest path between node *i* to node *j*. The BWC and CLC are determined by the Breadth-First Search algorithm (Cormen et al., 2009).

An adjacency matrix A is an n x n matrix where each cell a_{ij} = 1 if node i and node j are connected or a_{ij} = 0 if the nodes are not connected (n = number of nodes). Degree centrality (DGC) is the sum of contacts made by node i to other nodes or simply the number of neighbors that a node has. $DGC_i = \sum_j a_{ij}$, where a_{ij} is a link connecting node i to node j, with a value of 1 (connected) or 0 (not connected). A satellite node having only one connected neighbor has DGC = 1, while a central node connected to 10 other nodes has DGC = 10. Thus, nodes with high DGC values are likely to play a critical role in an epidemic. Eigenvector centrality (EVC) measures a node's degree as well as its neighbors' degrees. It measures the influence of a node in a network, i.e., a node is considered to be important if it is linked to other important nodes. EVC for a node i is the i-th element of an eigenvector x defined by $Ax = \lambda x$ where A is the adjacency matrix and λ is the eigenvalue i.e., $EVC_i = \frac{1}{\lambda}\sum_j a_{ij}x_j$. The EVC is determined by the power iteration algorithm (Chung, 2006).

The four centrality measures were calculated for each static network created for different τ values (described in section 2.2.2). We also chose one network where each node was connected to at least another node for further analysis. First, the empirical cumulative probability distributions of BWC, CLC, DGC, and EVC were calculated for each epidemic year to show the distribution of the calculated centrality values. For a set of values across a set of nodes, we calculated the probability of each value and used the empirical cumulative density function in the *ggpubr* package to calculate the cumulative probability distributions of BWC, CLC, DGC, and EVC. Second, the similarity in the ranking of nodes between BWC and all other centrality metrics was assessed using Spearman's rank-based correlation (a third analysis is described in section 2.3).

2.2.2. Identification of important nodes for disease spread within the static network

An analysis of the 2008 to 2016 observed disease occurrence data was necessary to determine if recurring patterns could help design effective disease control measures. A simple approach was to check for the nodes that reoccurred from year to year, i.e., nodes observed at least once. For example, one node was observed in 2008 and 2009 but not in the other years. This node reoccurred twice. Another node observed in 2008, 2009, 2010, 2011, 2013, and 2014 but not in 2015 and 2016, reoccurred six times. The number of times a node reoccurred from 2008 to 2016 was defined as the *infection frequency*. Thus, two approaches used to identify nodes that were important for disease spread and thus could be useful for risk-based surveillance and disease management to reduce epidemic spread: i) selection of nodes based on this infection frequency alone, and ii) selection of nodes based on a combination of this infection frequency and BWC, CLC, DGC, and EVC metrics. For option (i), the infection frequency was calculated for all nodes in the 2008 to 2016 dataset. The nodes were then were ranked from the highest to the lowest frequency value (Figure 3.2).

For option (ii), the 2008-2016 data set was reduced to contain only 2008-2016 nodes with infection frequency ≥ 1. A static network was created such that each node was connected to at least another node (Ferrari et al., 2014) using $b = 2.11$ (Ojiambo et al., 2017). BWC, CLC, DGC, and EVC centralities were calculated for this network. Then the BWC, CLC, DGC, and EVC metrics were scaled between 0 and 1 and combined in a ratio of 80:20 (frequency: centrality) for each node as described by Sutrave et al. (2012) to give more weight to the infection frequency (Figure 3.6). The nodes were then ranked in decreasing order based on this weighted value. This approach emphasizes nodes where the epidemic is active and highly connected nodes that act as bridges to

other nodes (for BWC), nodes that occur in the shortest path (for CLC), or nodes that are connected to other potential superspreaders (for DGC and EVC).

For each year, we considered threshold values for τ such that the bounds for τ produced dense networks and sparse networks, i.e., $10^{-18} < \tau < 10^{-8}$. A network was created for each τ, totaling 20 networks (with varying levels of network density as in Chapter 2) for all τ values. DGC centrality was calculated for each network, and the results were ranked in decreasing order, totaling 20 rankings for the 20 networks. The top 20 nodes with the highest scores were then selected (because the DGC ranking produced slightly different results for each of the 20 networks). A second ranking was done for each node in this top 20 set. The number of times a node appeared in the top 20 list across all thresholds was counted to eliminate the nodes that were ranked with higher scores in the dense and sparse networks. The nodes were then ranked in decreasing order. This process was repeated for the other three centrality measures. The results across the four measures and τ values (different networks) were combined in a heatmap using ggplot2 (Wickham, 2016) package in the R environment.

2.3. Dynamics on the static networks

Here, we describe the dynamic process, CDM spread, occurring on a static network where each node is connected to at least another node (Ferrari et al., 2014). Within a year, we modeled the probabilities of different nodes being infected over discrete weekly time steps $t \in \{1, 2 \ldots T\}$ based on the SI model described by Sutrave et al. (2012). We assumed that *P. cubensis* is mainly dispersed by wind, homogeneity in the host response to *P. cubensis,* and favorable weather for infection and CDM spread. The model combines the static components that are constant during

the cropping season and the dynamic components that vary during the cropping season and are formulated as

$$\begin{cases} \alpha_{ij} = \left(l_{ij}\right)^{-b} \\ \beta_{ij} = \frac{\vec{l_{ij}}.\vec{w_t}}{\left|\vec{l_{ij}}\right|} \\ u_{ij} = \alpha_{ij} \times \beta_{ij} \end{cases} \tag{3}$$

where α_{ij} is a constant function of the between-node distance and decays exponentially with distance, l_{ij} the distance between node *i* and node *j*, *b* is the spread parameter (Ojiambo et al., 2017), β_{ij} is a function of the wind data defined as the dynamic wind-based infection rate (Sutrave et al., 2012), $\vec{l_{ij}}$ is the displacement vector between two nodes, $\vec{w_t}$ is the wind vector at time *t*, and u_{ij} is the link weight based on distance and wind between node *i* and node *j* at time *t*.

Since the probability of a node being infected depends on the number of infected neighbors, the probability ϑ_i of node *i* not being infected by its neighbors was expressed as

$$\vartheta_i(t) = \prod_{j \in N_i} \left(1 - u_{ij} p_j(t)\right) \tag{4}$$

where p_j is the probability of node *j* being infected at time *t*, $u_{ij} \in [0,1]$ is the link weight as defined above, and N_i is a set of node i's neighbors. Given Equation 4, the probability p_i of node *i* being infected at time *t* was expressed as

$$p_i(t) = 1 - \left(1 - p_i(t-1)\right)\vartheta_i(t) \tag{5}$$

Values of β_{ij} were updated at each time step and p_i was calculated at each time step. In this study, a weekly time step was adopted, and all calculations were performed using MATLAB (R2019a).

2.3.1. Error quantification

A value of 1 was assigned to nodes observed as infected, while a value of 0 was assigned to healthy nodes in the observed data at each time step *t*. The error was defined as the absolute difference between the observed data and the corresponding infection probability calculated by the model at each time step *t*. As described by Sutrave et al. (2012), the mean error for the infected nodes at time step *t* is:

$$\hat{E}_{in}(t) = \frac{\sum_{i=1}^{N_{in}(t)}(1 - p_i(t))}{N_{in}(t)} \tag{6}$$

where $N_{in}(t)$ is the total number of infected nodes at time step *t*, while $p_i(t)$ is defined above. Similarly, the mean error for healthy nodes for time step *t* was calculated as:

$$\hat{E}_{hn}(t) = \frac{\sum_{i=1}^{N_{hn}(t)} p_i(t)}{N_{hn}(t)} \tag{7}$$

where $N_{hn}(t)$ is the total number of healthy nodes at time step t. The total error was obtained by using

$$\hat{E} = \alpha\hat{E}_{in}(t) + (1 - \alpha)\hat{E}_{hn}(t) \tag{8}$$

where the ratio of α: $(1 - \alpha)$ is 8:2 (α is a weighting factor) i.e., the observed-infected nodes were given four times more weight than the observed-healthy nodes in evaluating the final error (Sutrave et al., 2012).

2.3.2. Assessing node importance in a dynamic network based on centrality measures

The importance of nodes identified as highly connected, based on centrality measures from the static network analysis, was investigated for their impact on disease spread on the dynamic network. Nodes identified as most important based on BWC, CLC, DGC, and EVC values were removed from the networks, and the probabilities of disease spread among the remaining nodes

were recalculated in the new dynamic network for each epidemic year. Prediction of disease outbreaks based on all nodes in the network was subsequently compared to predictions where nodes identified as highly connected based on BWC, CLC, DGD, and EVC had been removed from the network. This approach is equivalent to intensive disease management where nodes are completely removed, and their resultant impact on disease propagation is assessed.

3. Results

3.1. Spatiotemporal dynamics of disease spread in the eastern United States

The observations of disease occurrences suggested a spatial association between the locations of first and last disease reports. CDM was first observed in a sentinel plot in southern Florida in Miami-Dade County in 6 out of 8 epidemic years (Figure 3.1). Across all years, the earliest date of the CDM epidemic in southern Florida was in February 2008. Most of the first CDM reports from 2009 to 2016 were in southern Florida in February and March. These epidemics were in both sentinel and non-sentinel plots (i.e., commercial and research plots). Although early reports were expected from the monitored sentinel plots, early reports were also made from non-sentinel plots before sentinel plots in some years. This is one reason we combined both sentinel and non-sentinel plots in this study.

Subsequent new CDM reports progressed northward with time. The new reports were made later in more northern states relative to reports in the southern states (Figure 3.1). In 2009, CDM was first reported in a commercial plot in southwestern Texas along the Gulf of Mexico around the same time the disease was reported in northern Florida and southern Georgia (Figure 3.1). The first CDM reports in northern states (e.g., Michigan, New York, or Wisconsin) occurred considerably later than corresponding reports of first CDM outbreaks in southern states (e.g.,

Alabama, Georgia, or South Carolina) (Figure 3.1). The last set of new disease reports across all years were in Indiana, Illinois, Kentucky, Louisiana, Massachusetts, Maryland, Michigan, Minnesota, Mississippi, Missouri, New Jersey, New York, North Carolina, Ohio, Pennsylvania, South Carolina, Tennessee, Texas, Vermont, Virginia, West Virginia, and Wisconsin, in July, August, and September.

The total number of states with CDM ranged from 22 to 27. The corresponding number of counties ranged from 86 to 179 across the region (Table 3.1). There was a positive correlation between the number of disease reports and counties; R = 0.95, p = 0.00068 (Figure S3.2), i.e., the number of plots increased as the number of infected counties increased. The correlation between the number of counties where the disease is present vs. the number of counties where surveillance was occurring was R = 0.37, p = 0.33 (Figure S3.3). The maximum distance between two CDM reports, a measure of epidemic extent, ranged from 2,491 km in 2012 (that was between a research plot in Texas and a commercial field in Massachusetts) to 3,071 km in 2015 (which was between a commercial field in Texas and a commercial field in New Hampshire).

3.2. Infection based frequency selection of important nodes

The number of times nodes were infected based on combined epidemic data in all epidemic years varied from 1 to 6 (Figure 3.2). Nodes where the infection frequency was consistently higher (i.e., frequency > 3) were in Alabama, Maryland, Michigan, North Carolina, Ohio, and South Carolina. The nodes with the highest infection frequency were in Wicomico County in Maryland, Johnson, Lenoir, New Hanover, and Sampson counties in North Carolina, Sandusky, Huron, and Wayne counties in Ohio with infection frequency of 5 and 6 (Figure 3.2). The remaining nodes

with a low infection frequency (i.e., frequency ≤ 3) constituted most of the nodes in counties scattered throughout the study region.

3.3. Centrality measures and selection of important nodes

The betweenness, closeness, degree, and eigenvector values varied between epidemic years. Further, variability in individual metrics was also observed within a given epidemic year for a chosen static network where a node was connected to at least another node. We looked at the standard deviation of each measure to check for variability within the measure, not across the measures. Betweenness centrality of nodes within a network was the most variable within any epidemic year across the entire study. For example, BWC values ranged from 264.5 to 888.3 in 2008 (Table 3.3), from 1147.6 to 2415.7 in 2009 (Table 3.4), and from 237.6 to 1718.2 in 2010 (Table 3.5). The mean BWC for the twenty most important nodes in these respective years were 441.8, 1656.9, and 474, with corresponding standard deviation values of 441.1, 896.7 and 1046.9, in respective years. These BWC values are the number of times a node lies on the shortest path considering all pairs of nodes in the network. An example explaining BWC is given in the appendix. The standard deviations of closeness, degree, and eigenvector values were comparatively less variable among nodes than BWC (Tables 3.3, 3.4, and 3.5), with CLC being the least variable across the entire study.

Cumulative probability distribution of BWC values for the nodes in the networks considered exhibited a power-law style distribution. About 85% of the nodes have BWC values less than 250, with BWC > 1500 being the largest BWC value observed (Figure 3.3). In contrast, the cumulative distribution of CLC values was uniformly distributed within a smaller range, resulting in a relatively steep curve (Figure 3.3). The cumulative distribution of EVC values

followed a Poisson distribution (Figure 3.3). Apart from the most important node in each year (EVC = 1), each node has an EVC value closer to one or two other nodes. These were the same features observed by Meghanathan and Lawrence, 2016.

The ranking of important nodes varied across centrality measures and epidemic years examined in this study. For example, the node ranked as the most important in 2008 based on BWC was node 74 in Mississippi. While node 89 in Surry County in North Carolina, node 131 in Centre County, and node 128 Westmoreland County in Pennsylvania, were ranked as the most important based on CLC, DGC, and EVC, respectively (Table 3.3). Similarly, node 34 in Spalding County in Georgia was ranked the most important in 2009 based on BWC. In contrast, nodes 121 in Bertie County in North Carolina, 74 in Lenawee County in Michigan, and 109 in Franklin County in North Carolina were the most important based on CLC, DGC, and EVC, respectively (Table 3.4). In 2010, node 30 in Harrison County in Kentucky was ranked as the most important based on BWC and CLC, while node 116 in Summit County in Ohio was ranked the most based on DGC and EVC (Table 3.5). The ranking of nodes based on BWC and other centrality metrics varied across epidemic years. In general, Spearman's rank-based correlation coefficients were highest between BWC and CLC, with correlations ranging from 0.43 to 0.74 (Figure 3.4). Correlations between BWC and DGC or EVC were relatively lower across the epidemic years except between BWC and DGC in 2016, where $r = 0.46$ (Figure 3.4).

The consistency in the rankings of nodes based on BWC, CLC, DGC, and EVC was summarized as a heatmap (Figure 3.5) to visualize unique nodes within the networks. For example, many nodes overlapped in their rankings among the top 20 important nodes (across all thresholds and centralities) in 2010 (Figure 3.5A) and 2014 (Figure 3.5C) based on BWC and CLC. However, most nodes overlapped across all the four centrality measures in 2011 (Figure 3.5B). For example,

node 117 in Lewis County in West Virginia appeared more than 20 times in the top 20 ranks based on BWC and CLC. This same node also appeared more than ten times in the top 20 ranks based on DGC and EVC.

3.4. Infection frequency and centrality selection of important nodes

Identifying important nodes based on infection frequency and centrality (from static networks) showed some similarities and differences based on the examined centrality metric. The ranking of nodes based on BWC and CLC was generally similar across years, while rankings based on EVC were different from all other centrality measures. Based on BWC, nodes that had frequency > 4 had the highest calculated values (combined frequency x centrality), with the largest value being 0.82 for the node in Sandusky County in Ohio (Figure 3.6), while the lowest weight was 0.13 for a node in Charleston County in South Carolina. Based on CLC, the largest weight for the source was 0.98 for the node in Sandusky County in Ohio that had a frequency > 6, with the node with the lowest weight being a node in Miami-Dade County in Florida with a weight of 0.198. Similarly, the node in Sandusky County in Ohio had the highest weight of 0.93 based on DGC, followed by nodes in Johnston, Lenoir, New Hanover counties in North Carolina, Wicomico County in Maryland, and Huron and Wayne counties in Ohio that has a frequency of 5 (Figure 3.6). Node ranking based on EVC was comparably different from a ranking based on all other centrality measures. A node in Johnston County in North Carolina had the highest weight of 0.84, followed by nodes Wicomico County in Maryland, Sampson, and Johnston counties North Carolina and Wayne County in Ohio (Figure 3.6). The node with the lowest was the same as that identified by CLC.

3.5. Dynamic network model for cucurbit downy mildew

The dynamic network model revealed an evolving CDM network where the probability for a node to be infected increased in time (Figure 3.7). In 2014, nodes closest to the initial disease source in Miami-Dade County in Florida (open square) had a probability of 1 of getting infected early in the season, i.e., the nodes in Florida and Georgia closest to the initial disease occurrence in southern Florida had the highest probabilities by week 10. In contrast, the probability of infection for nodes in the rest of the network was 0 (Figure 3.7). As time progressed, the probabilities increased for nodes that were further away from the source node. However, probabilities remained at 0 for isolated nodes as disease spread proceeded in time and space (Figure 3.7). As the epidemic advanced at week 15, the probabilities of infection for nodes in Georgia increased, while the probability of infection for nodes elsewhere in the northeast U.S. was 0.

At week 20, the probability for infection of nodes in South Carolina increased noticeably. In contrast, probabilities were low for nodes in more northern states (e.g., North Carolina and Tennessee) or in states where no disease was reported. At week 25, the probabilities for node infection increased for nodes in North Carolina, and at week 30, the probabilities increased for other nodes in North Carolina, Virginia, New York, and Pennsylvania (Figure 3.7). At week 35, the probabilities increased for other nodes in the eastern United States, with only a few nodes in Illinois and Michigan having low infection probabilities. Similar patterns of infection probabilities were observed in other epidemic years, except that the strength of these probabilities differed between years (see supplementary materials).

3.6. Errors in dynamic model and node importance based on centrality measures

The means of the absolute errors generated across weekly time steps and averaged monthly from January to August between the observed and predicted, healthy and infected nodes in the network varied depending on the epidemic year used to construct the dynamic networks. For example, the lowest mean absolute error was 0.099, which was observed for the network in 2016, while the highest mean absolute error was 0.353, which was observed in the network of 2010 (Table 3.6). Low errors comparable to those observed in 2016 were also observed for epidemic data in 2014 and 2015, where the absolute errors were 0.168 and 0.121, respectively. The mean absolute error across all epidemic years was low, with an error of 0.228 (Table 3.6).

Removal of nodes identified as important based on BWC, CLC, DGC, and EVC impacted errors of the dynamic model across epidemic years, i.e., removal of nodes resulted in less accurate predictions of the spread that was observed (Table 3.6). The impact on model performance (i.e., increase in error) when nodes (identified to be important based on centrality measures) were removed from the network was assessed. Removal of nodes identified as important by BWC resulted in an error rate of 0.338 (Table 3.6), representing a 33.8% error rate relative to the base prediction that had an error of 0.228. In contrast, removing nodes identified as important based on CLC, DGC, and EVC marginally increased model errors to 0.248, 0.256, and 0.261, respectively. These error rates represented an 8.8, 12.2, and 14.5%, respectively, increase in error relative to the base prediction with all the nodes present in the network. In addition, removing nodes identified as important based on BWC resulted in 3 to 6 times higher (Table 3.6) than errors resulting from the removal of nodes identified as important based on CLC, DGG, and EVC.

The probability of node infection and epidemic progress in the network was also differentially impacted by removing nodes identified as important based on centrality measures

examined. Relative to a network with all nodes present, removing nodes identified as important based on BWC reduced the probability of infection of uninfected nodes in the subsequent time step in all epidemic years (Figure 3.8). For example, the removal of the nodes in north Florida, Georgia, and South Carolina that were identified as important based on BWC did not allow the progression of CDM and infection of nodes in north Florida, South Georgia, and South Carolina in 2009 by week 25 (Figure 3.8). Similarly, when nodes were identified to be important based on BWC were removed in Alabama and Georgia in 2014, the infection probability of nodes in South Carolina and North Carolina was greatly reduced. A similar pattern of the probabilities was based on BWC was observed in other epidemic years except that the magnitude of these probabilities differed between years (see supplementary materials). However, removing nodes identified as important based on either CLC, DGC, or EVC had a minor impact and progression of the epidemic in the subsequent time step in all epidemic years (Fig 3.8; see supplementary materials).

4. Discussion

In this study, networks based on historical epidemic records collected from 2008 to 2016 were formulated to describe how the dispersal of *Pseudoperonospora cubensis* and the spread of cucurbit downy mildew from infected to disease-free cucurbit fields are influenced by the connectivity of cucurbit fields. We began by analyzing multiple low to high-density static networks and chose individual networks for each year for further analysis. This is because high-density networks (generated at low thresholds) have shorter between-node path lengths for a pathogen to travel and more dispersal pathways, while a pathogen has fewer shorter links and dispersal pathways in low-density networks (generated at high threshlolds) (Ames et al., 2011). Therefore, individual networks where a node was connected to at least another node (Ferrari et al.,

2014) were selected for further analysis because their intermediate density structures are assumed to impact disease behavior (Ames et al., 2011), wherein disease dynamics can be explained by network analysis (Christley et al., 2005; Wang 2003; Xing et al., 2020).

At the center of the CDM ipmPIPE surveillance platform is a series of sentinel plots in the eastern United States monitored for disease outbreaks by state collaborators. The surveillance system has helped predict when to apply the first fungicide spray against cucurbit downy mildew (Neufeld et al., 2018). The goal for a surveillance system is early disease detection and to document the absence of disease (Martin et al., 2007). However, like many other disease surveillance systems, the CDM ipmPIPE system is expensive to maintain, and resources are often limited. Thus, targeted sampling of highly connected sites critical in spreading the disease may undoubtedly benefit disease surveillance. Network centrality measures such as betweenness (BWC), closeness (CLC), degree (DGC), and eigenvector (EVC) can help identify highly connected nodes (Andersen et al., 2019; Gent et al., 2019; Sankara et al., 2015) and evaluate strategies for selecting plots for surveillance under different scenarios of resource limitation (Sankara et al., 2015). These centrality measures have implications in disease epidemics. Node importance is determined by either the number of connections the node has (i.e., DGC), the number of connections the node's neighbors have (i.e., EVC), how a node acts as a bridge to other nodes (i.e., BWC), or the short average distance from that node to all other nodes (i.e., CLC). Thus, the disease will spread faster and with a high probability from nodes with a high BWC, while a disease spreading from a node with high CLC would reach all other connected nodes in a shorter number of steps.

Based on a complete static network model, the four centrality measures were used to identify the highly connected cucurbit fields that may have important implications for surveillance and controlling CDM. The importance of these highly connected fields in disease control was also

evaluated using a dynamic network model. In this study, BWC was more useful in identifying the important cucurbit fields that could be targeted for surveillance and disease control to reduce epidemic spread. Nodes identified as important based on either CLC, DGC, or EVC did not reduce the probability of node infection in subsequent time steps compared to a scenario where all the nodes were present in the network. Further, removing these nodes did not affect errors in the dynamic model, unlike removing nodes identified as important based on BWC. Across epidemic years, more central nodes identified as important based on BWC were in Michigan in the Great Lakes region, Ohio in the mid-west, and Maryland, North Carolina, South Carolina, and Virginia along the Atlantic coast. Thus, these nodes could be reasonable targets for more intensive sampling for surveillance and management to reduce inoculum production that drives infection in neighboring cucurbit fields in the eastern United States.

The spatial location and connectivity of nodes in the networks influenced the node removal analysis. For example, for 2008, the top-ranked nodes were located in Pennsylvania, Ohio, and New York, based on DGC and EVC rankings. Removing these nodes did not affect disease progression in the southern states. However, for 2014, the top-ranked nodes were located in North Carolina based on DGC and EVC rankings. Removing these nodes affected the disease spread in the southern and northern states. Also, for 2008, 2009, and 2013, top-ranked nodes based on CLC were located mainly in the central-eastern US, and removing them did not stop disease spread. However, in 2014 top-ranked nodes were located mainly in North Carolina, and removing them stopped disease spread past North Carolina. However, nodes with high BWC scores were scattered all over the map, and their removal broke the networks. For example, for 2008 and 2014, the top-ranked nodes were located in Georgia, and removing them stopped the disease from spreading past

Georgia. In 2009 and 2013, the top-ranked nodes were located in Florida. Removing these nodes stopped the disease from spreading past Florida.

The addition of centrality measures to the frequency of node infection substantially improved the identification of important nodes. For example, DGC, BWC, and CLC produced similar rankings with the infection-based frequency for nodes with an infection frequency > 4. Although EVC produced a different ranking, nodes with frequency > 4 still had high weights, thus agreeing with the rankings from the other centrality measures. The combination of frequency-based and DGC was useful in selecting sampling nodes for sentinel plots for soybean rust in the United States (Sutrave et al., 2012). The DGC is a standard measure in network science and is useful for identifying important nodes in static networks of several pathosystems to inform strategic management (Christley et al., 2005; Gent et al., 2019; Kiss et al., 2006; Xing et al., 2020). Unlike other centrality measures, DGC is easier to calculate and does not require assessing the entire network (Christely et al., 2005). In this study, DGC was not effective in identifying important nodes as compared to BWC. Further, BWC rankings were poorly correlated with those of DGC except in epidemic data collected in 2016. The latter is an indicator of greater variation in the ability of different centrality measures to predict the risk of CDM outbreaks.

The analysis over multiple thresholds presented in this study demonstrates that the characteristics of a 'scale-free network depend on the cut-off value used to generate a network. Scale-free networks are networks characterized by large hubs, i.e., a network with a power-law degree distribution. A few networks for 2008, 2011, and 2014 are scale-free (Fig S3.19). These networks are characterized by many nodes with few links and a few nodes with many links resulting in a right-skewed distribution for the number of links that follow a power-law distribution (Banks et al., 2015; Barabási and Albert 1999). Pathogens can disperse easily and quickly in scale-

free networks via highly connected nodes containing many links (Jeger et al., 2007). Further, an infection threshold is absent in scale-free networks (Pastor-Satorras and Vespignani 2001), and all infections can result in an epidemic making these networks very vulnerable.

This study was conducted to investigate the dispersal of *P. cubensis* and the spread of CDM in simple networks to draw broad conclusions on the utility of centrality measures in predicting the probability of infection to inform surveillance and management of the disease. This study will add efficiency to the current framework for predicting the initial outbreak of the disease in the eastern United States. One basic approach to limiting disease spread at the landscape level is applying targeted treatment to specific fields within the affected area. Estimating the probability and timing of disease outbreaks in specific locations and determining where and when the introduction of inoculum can impact the extent of an epidemic is one of the challenges in predicting disease spread (Meentemeyer et al., 2011; Fitzpatrick et al., 2012). Thus, the ability to rapidly identify these fields within the network and mobilize the necessary resources is key to successful mitigation. Locations identified as highly connected in the network can be targeted for early surveillance when collecting reports of new diseases within the region. These locations can also be targeted for fungicide treatment to slow down the rate of inoculum production and dispersal to disease-free neighboring cucurbit fields. Thus, centrality measures provide a greater understanding of infection dynamics to inform surveillance and management of CDM. Degree centrality, which is more readily measured, was not as good as other measures such as betweenness centrality in identifying highly connected nodes and predicting risk of infection in the CDM network.

Unlike the dynamic model used for soybean rust spread in the United States, the dynamic model used in this study incorporated a power-law dispersal gradient characteristic for long-distance dispersal of plant pathogens. Based on 2008 and 2009 data and bivariate O-ring statistics,

Ojiambo et al. (2011) found that the spatial spread of CDM cases was 0 to 390 km, 737 km, 879 km, with 1000 km being the maximum possible distance. Further, Ojiambo et al. (2017) showed that the spread parameter *b* is unstable, with the final epidemic extent ranged from 4.16×10^8 km^2 to 6.44×10^8 km^2. This is why the analysis was done using different *b* values to account for the difference in spatial spread. This model improves on long-distance dispersal by using a flexible threshold for the distance to allow for connectivity of further apart nodes. However, the model does not account for differences in environmental factors that are likely to influence pathogen dispersal. In addition, accounting for differences in host susceptibility at the different locations could further improve our ability to generalize the findings reported here to different cucurbit host types.

References

1. Ames, G. M., George, D. B., Hampson, C. P., Kanarek, A. R., McBee, C. D., et al., 2011. Using network properties to predict disease dynamics on human contact networks. Proc. R. Soc. B 278:3544-3550.
2. Andersen, K. F., Buddenhagen, C. E., Rachkara, P., Gibson, R., Kalule, S., Phillips, D., and Garrett, K. A. 2019. Modeling epidemics in seed systems and landscapes to guide management strategies: The case of sweet potato in northern Uganda. Phytopathology 109:1519-1532.
3. Banks, N. C., Paini, D. R., Bayliss, K. L., and Hodda, M. 2015. The role of global trade and transport network topology in the human-mediated dispersal of alien species. Ecol. Lett. 18:188-199.
4. Barabási, A. -L. and Albert, R. 1999. Emergence of scaling in random networks. Science 286:509-512.
5. Bigras-Poulin, M., Barfod, K., Mortensen, S., and Greiner, M., 2007. Relationship of trade patterns of the Danish swine industry animal movements network to potential disease spread. Prevent. Vet. Med. 80: 143-165.
6. Brown, J. K., and Hovmøller, M. S. 2002. Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. Science 297:537-541.
7. Campbell, C. L., and Madden, L. V. 1990. Introduction to Plant Disease Epidemiology. Wiley, New York.
8. Choi, Y. J., Hong, S. B., and Shin, H. D. 2005. A re-consideration of *Pseudoperonospora cubensis* and *P. humuli* based on molecular and morphological data. Mycol. Res. 109:841-848.

9. Chung, F, Complex Graphs and Networks, American Mathematical Society, 1st edition, (2006).
10. Christley, R. M., Pinchbeck, G. L., Bowers, R. G., Clancy, D., Frnech, N. P., Bennett, R., and Turner, J. 2005. Infection in social networks: Using network analysis to identify high-risk individuals. Am. J. Epidemiol. 162:1042-1031.
11. Cohen, Y., van den Langenberg, K. M., Wehner, T. C., Ojiambo, P. S. Hausbeck, M., Quesada-Ocampo, L. M., Lebeda, A., Sierotzki, H., and Gisi, U. 2015. Resurgence of *Pseudoperonospora cubensis*: The causal agent of cucurbit downy mildew. Phytopathology 105: 998-1012.
12. Cormen, T.H, Leiserson, C.E, Rivest, R.L, and Stein, C. Introduction to algorithms, MIT Press, 3rd Edition (2009).
13. Crowl, T. A., Crist, T. O., Parmenter, R. R., Belovsky, G., and Lugo, A. E. 2008. The spread of invasive species and infectious disease as drivers of ecosystem change. Front. Ecol. Environ. 6:238-246.
14. Csardi, G., and Nepusz, T. 2006. The igraph software package for complex network research. InterJournal, Complex Systems, 1695. http://igraph.org.
15. Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V, and Vernon, M. C. 2010. Networks and the epidemiology of infectious disease. Interdiscip. Perspect. Infect. Dis. Volume 2011, Article ID 284909.
16. Farine, D. R., and Whitehead, H. 2015. Constructing, conducting and interpreting animal social network analysis. J. Anim. Ecol. 84:1144-1163.
17. Ferguson, N. M., Donnelly, C. A., Anderson, R. M. 2001. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. Science 292: 1155-1160.

18. Ferrari, J. R., Preisser, E. L., and Fitzpatrick, M. C. 2014. Modeling the spread of invasive species using dynamic network models. Biol. Invasions 16: 949-960.
19. Firester, B., Shtienberg, D., and Blank, L. 2018. Modelling the spatiotemporal dynamics of Phytophthora infestans at a regional scale. Plant Pathol. 67: 1552-1561.
20. Garrett, K. A., Alcal´a-Brisen˜o, R. I., Anderson, K. F., Buddenhagen, C. E., Choudhury, R. A., Fulton, J. C., Hernandez Nopsa, J. F., Poudel, R., and Xing, Y. 2018. Network analysis: A systems framework to address grand challenges in plant pathology. Annu. Rev. Phytopathol. 56:25.1-25.22.
21. Gent, D. H., Bhattacharyya, S., and Ruiz, T. 2019. Prediction of spread and regional development of hop powdery mildew: A network analysis. Phytopathology 109:1392-1403.
22. Hernandez Nopsa, J. F., Daglish, G. J., Hagstrum, D. W., Leslie, J. F., Phillips, T. W, Scoglio, C, Thomas-Sharma, S., Walter, G. H., and Garrett, K. A. 2015. Ecological networks in stored grain: key postharvest nodes for emerging pests, pathogens, and mycotoxins. Bioscience 65:985-1002.
23. Hijmans, R. J. 2017. geosphere: Spherical Trigonometry. R Package Version 1.5-7. https://cran.r-project.org/web/packages/geosphere/index.html
24. Holme, P. 2018. Objective measures for sentinel surveillance in network epidemiology. Phys. Rev. E 98:022313.
25. Holme, P. 2017. Three faces of node importance in network epidemiology: Exact results Picture for small graphs. Phys. Rev. E 96:062305.
26. Jeger, M. J. 1999. Improved understanding of dispersal in crop pest and disease management: current status and future directions. Agric. For. Meteorol. 97: 331-349.

27. Jeger, M. J., Pautasso, M., Holdenrieder, O., Shaw, M. W. 2007. Modelling disease spread and control in networks: implications for plant sciences. New Phytol. 174: 279-297.
28. Kao, R. R., Danon, L., Green, D. M., Kiss, I. Z. 2006. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. Proc. Royal Soc. B: 273:1999-2007.
29. Keeling, M. J., and Eames, K. T. 2005. Networks and epidemic models. J. Roy. Soc. Interface 2: 295-307.
30. Kiss, I. Z., Green, D. M., and Kao, R. R. 2006. The network of sheep movements within Great Britain: Network properties and their implications for infectious disease spread. J. R. Soc. Interface 3:669-677.
31. Lamour, A., Termorshuizen. A. J., Volker, D., and Jeger, M. J. 2007. Network formation by rhizomorphs of *Armillaria lutea* in natural soil: their description and ecological significance. FEMS Microbiol. Ecol. 62:222-232.
32. Lookingbill, T. R., Gardner, R. H., Ferrari, J. R., and Keller, C. E. 2010. Combining a dispersal model with network theory to assess habitat connectivity. Ecol. Appl. 20:427-441.
33. Main, C. E., Keever, T., Holmes, G. J., and Davis, J. M. 2001. Forecasting long-range transport of downy mildew spores and plant disease epidemics. APSnetFeature-2001-0501
34. Meghanathan, N and Lawrence, R Centrality analysis of the United States network graph, 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS 2016), 2016, pp. 1-6

35. Meentemeyer, R. K., Cunniffe, N. J., Cook, A. R., Joao, J. A., Hunter, R. D., Rizzo, D. M., and Gilligan, C. A. 2011. Epidemiological modeling of invasion in heterogeneous landscapes: Spread of sudden oak death in California (1990-2030). Ecosphere 2(2).
36. Moslonka-Lefebvre, M., Finley, A., Dorigatti, I., Dehnen-Schmutz, K., Harwood, T., Jeger, M. J., Xu, X., Holdenrieder, O., and Pautasso, M. 2011. Networks in plant epidemiology: From genes to landscapes, countries, and continents. Phytopathology 101:392-403.
37. Mundt, C. C., Sackett, K. E.,Wallace, L. D., Cowger, C., and Dudley, J. P. 2009. Long-distance dispersal and accelerating waves of disease: empirical relationships. Am. Nat. 173:456-466.
38. Natale, F., Giovannini, A., Savini, L., Plama, D., Possenti, L., Fiore, G., and Calistri, P. 2009. Network analysis of Italian cattle trade patterns and evaluation of risks for potential disease spread. Prevent. Vet. Med. 92: 341-350.
39. Neufeld, K. N., Keinath, A. P., Gugino, B. K., McGrath, M. T., Sikora, E. J., Miller, S. A., Ivey, M. L., Langston, D. B., Dutta, B., Keever, T., Sims, A., and Ojiambo, P. S. 2018. Predicting the risk of cucurbit downy mildew in the eastern United States using an integrated aerobiological model. Int. J. Biometeorol. 62:655-668.
40. Ojiambo, P. S., and Holmes, G. J. 2011. Spatiotemporal spread of cucurbit downy mildew in the eastern United States. Phytopathology 101:451-461.
41. Ojiambo, P. S., Gent, D. H., Mehra, L. K., Christie, D., and Magarey, R. 2017. Focus expansion and stability of the spread parameter estimate of the power law model for dispersal gradients. PeerJ 5:e3465

42. Ojiambo, P. S., Holmes, G. J., Britton, W., Keever, T., Adams, M. L., Babadoost, M., Bost, S. C., Boyles, R., Brooks, M., Damicone, J., Draper, M. A., Egel, D. S., Everts, K. L., Ferrin, D. M., Gevens, A. J., Gugino, B. K., Hausbeck, M. K., Ingram, D. M., Isakeit, T., Keinath, A. P., Koike, S. T., Langston, D., McGrath, M. T., Miller, S. A., Mulrooney, R., Rideout, S., Roddy, E., Seebold, K. W., Sikora, E. J., Thornton, A., Wick, R. L., Wyenandt, C. A., and Zhang, S. 2011. Cucurbit downy mildew ipmPIPE: a next generation web-based interactive tool for disease management and extension outreach. Online. Plant Health Progress. Online publication. PHP-2011-0411-01-RV.

43. Ojiambo, P. S., Gent, D. H., Quesada-Ocampo, L. M., Hausbeck, M. K., and Holmes, G. J. 2015. Epidemiology and population biology of *Pseudoperonospora cubensis*: a model system for management of downy mildews. Annu. Rev. Phytopathol. 53: 223–246.

44. Oldham, S. I., Fulcher, B., Parkes, L., Arnatkevic, A., Suo, C., and Fornito, A. 2019. Consistency and differences between centrality measures across distinct classes of networks. PLoS ONE 14:e0220061.

45. Pastor-Satorras R., and Vespignani, A. 2001. Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86:3200-3203.

46. Pautasso M., and Jeger, M. J. 2008. Epidemic threshold and network structure: The interplay of probability of transmission and of persistence in small-size directed networks. Ecol. Complex. 5:1-8.

47. R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

48. Sanatkar, M. R., Scoglio, C., Natarajan, B., Isard, S. A., and Garrett, K. A. 2015. History, epidemic evolution, and model burn-in for a network of annual invasion: Soybean rust. Phytopathology 105:947-955.

49. Sutrave, S., Scoglio, C., Isard, S. A., Hutchinson, J. M. S., and Garrett, K. A. 2012. Identifying highly connected counties compensates for resource limitations when evaluating national spread of an invasive pathogen. PLoS ONE 7:e37793.

50. Thomas, A., Carbone, I., Choe, K., Quesada-Ocampo, L. M., and Ojiambo, P. S. 2017. Resurgence of cucurbit downy mildew in the United States: Insights from comparative genomic analysis of *Pseudoperonospora cubensis*. Ecol. Evol. 7:6231-6246.

51. Thomas, C. E. 1996. Downy mildew. Pages 25-27 in: Compendium of Cucurbit Diseases. T. A. Zitter, D. L. Hopkins, and C. E. Thomas, eds. American Phytopathological Society Press, St. Paul, MN.

52. With, K. A., Gardner, R. H., and Turner, M. G. 1997. Landscape connectivity and population distributions in heterogeneous environments. Oikos 78:151-169.

53. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

54. Virginia Weather & Climate Data. (n.d.). http://colaweb.gmu.edu/dev/clim301/lectures/wind/wind-uv.

Tables

Table 3.1. The number of plots with disease summarized by planting type where cucurbit downy mildew was reported during the study period.

Year	Number of states affected	Number of counties	Number by planting type					Total
			Commercial	Home garden	Research	Sentinel[a]	Unspecified[b]	
2008	22	113	68	10	12	59	5	154
2009	24	165	77	26	24	92	1	220
2010	25	118	77	17	24	25	1	144
2011	23	86	57	10	22	28	0	117
2012	25	149	99	20	23	31	0	173
2013	26	179	118	30	23	29	4	204
2014	23	104	53	16	22	20	3	114
2015	27	171	126	15	22	42	4	209
2016	22	107	61	9	19	33	0	122

[a] Sentinel planting type refers to fixed plots, planted early and designated for weekly monitoring.
[b]Unspecified refers to reports where the planting type was not stated when disease was reported in cucurbit downy mildew monitoring database.

Table 3.2. Definition of centrality measures in a network model used to study the spread of cucurbit downy mildew in the eastern United States.

Centrality measure	Central node	Relevance in epidemic spread
Betweenness	Acts as a bridge to other nodes	Removal of nodes with high betweenness may contain an epidemic
Closeness	Lies on the shortest path	Nodes are able to spread disease through a network
Degree	Connected to many other nodes	Nodes with high degree may be ‘superspreaders’
Eigenvector	Connected to other other high-degree nodes	Nodes with neighbors having high degree may be ‘superspreaders’

Table 3.3. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2008.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	74	MS	888.3	89	NC	0.0034	131	PA	73	128	PA	1.000
2	118	OH	665.3	118	OH	0.0034	52	MD	72	131	PA	0.994
3	135	SC	608.6	125	PA	0.0034	125	PA	72	134	PA	0.989
4	124	OH	534.1	128	PA	0.0034	128	PA	72	125	PA	0.981
5	39	KY	517.2	130	PA	0.0034	130	PA	72	130	PA	0.974
6	141	TN	507.2	124	OH	0.0034	127	PA	71	99	NY	0.963
7	31	GA	500.4	52	MD	0.0034	134	PA	69	127	PA	0.962
8	89	NC	471.1	134	PA	0.0034	99	NY	66	102	NY	0.953
9	137	SC	470.8	86	NC	0.0033	102	NY	65	96	NY	0.943
10	82	NC	416.6	148	VA	0.0033	96	NY	64	97	NY	0.930
11	91	NC	416.6	150	VA	0.0033	129	PA	64	98	NY	0.926
12	139	TN	375.8	131	PA	0.0033	11	DE	63	100	NY	0.902
13	52	MD	372.1	87	NC	0.0033	97	NY	63	52	MD	0.879
14	75	MS	336.7	88	NC	0.0033	98	NY	63	126	PA	0.858
15	125	PA	324.7	127	PA	0.0033	13	DE	62	129	PA	0.856
16	128	PA	305.4	80	NC	0.0033	100	NY	61	111	OH	0.847
17	136	SC	290.5	78	NC	0.0033	10	DE	59	113	OH	0.847
18	33	GA	290.1	79	NC	0.0033	93	NJ	59	117	OH	0.828
19	29	GA	279.0	151	VA	0.0032	94	NJ	59	120	OH	0.820
20	34	GA	264.5	39	KY	0.0032	133	PA	59	101	NY	0.814
Mean			441.8			0.0033			65.4			0.913
SD			441.1			0.0000			9.9			0.132

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation.

Table 3.4. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2009.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	34	GA	2415.7	122	NC	0.0012	74	MI	35	109	NC	1.000
2	212	VA	2390.2	132	NC	0.0012	79	MI	35	136	NC	0.979
3	48	KY	2376.2	134	NC	0.0012	82	MI	33	114	NC	0.979
4	154	OH	2152.4	129	NC	0.0012	93	MI	33	118	NC	0.966
5	32	GA	2011.5	124	NC	0.0012	109	NC	33	130	NC	0.960
6	192	TN	1907.7	135	NC	0.0012	158	OH	33	127	NC	0.960
7	186	SC	1803.5	205	VA	0.0012	200	VA	33	211	VA	0.937
8	169	PA	1796.5	212	VA	0.0012	76	MI	32	119	NC	0.913
9	2	AL	1672.3	48	KY	0.0011	90	MI	32	128	NC	0.906
10	180	SC	1605.4	163	OH	0.0011	114	NC	32	207	VA	0.898
11	104	MS	1515.0	164	OH	0.0011	118	NC	32	115	NC	0.891
12	171	PA	1413.6	165	OH	0.0011	136	NC	32	125	NC	0.887
13	103	MS	1351.4	133	NC	0.0011	211	VA	32	126	NC	0.884
14	25	FL	1343.5	192	TN	0.0011	75	MI	31	113	NC	0.882
15	200	VA	1311.5	123	NC	0.0011	83	MI	31	121	NC	0.872
16	153	OH	1259.8	169	PA	0.0011	88	MI	31	120	NC	0.869
17	147	NY	1258.1	171	PA	0.0011	89	MI	31	112	NC	0.869
18	54	KY	1248.4	183	SC	0.0011	91	MI	31	203	VA	0.867
19	101	MS	1158.2	207	VA	0.0011	92	MI	31	200	VA	0.864
20	158	OH	1147.6	203	VA	0.0011	111	NC	31	110	NC	0.850
Mean			1656.9			0.0011			32.2			0.912
SD			896.7			0.0000			2.8			0.106

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation.

Table 3.5. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2010.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	30	KY	1718.2	30	KY	0.0033	116	OH	56	116	OH	1.000
2	31	KY	1009.3	31	KY	0.0032	103	OH	54	109	OH	0.998
3	65	MS	691.0	116	OH	0.0032	104	OH	54	106	OH	0.997
4	4	AL	577.1	121	PA	0.0032	105	OH	54	110	OH	0.995
5	139	TX	556.0	105	OH	0.0032	106	OH	54	103	OH	0.995
6	77	NC	486.1	103	OH	0.0032	108	OH	54	113	OH	0.995
7	25	GA	469.3	104	OH	0.0032	109	OH	54	114	OH	0.995
8	74	NC	410.1	108	OH	0.0032	110	OH	54	104	OH	0.995
9	13	FL	404.1	110	OH	0.0032	113	OH	54	108	OH	0.995
10	23	GA	342.0	113	OH	0.0032	114	OH	54	61	MI	0.992
11	26	GA	342.0	114	OH	0.0032	61	MI	53	41	MI	0.983
12	5	AL	331.3	107	OH	0.0032	40	MI	52	53	MI	0.983
13	120	PA	305.2	106	OH	0.0031	41	MI	52	60	MI	0.983
14	130	SC	296.8	109	OH	0.0031	48	MI	52	48	MI	0.983
15	138	TX	282.0	120	PA	0.0031	53	MI	52	105	OH	0.977
16	80	NC	264.0	119	PA	0.0031	60	MI	52	42	MI	0.964
17	67	NC	257.1	115	OH	0.0031	107	OH	52	40	MI	0.960
18	117	PA	253.7	61	MI	0.0031	112	OH	52	111	OH	0.960
19	122	PA	246.7	112	OH	0.0031	122	PA	52	112	OH	0.959
20	140	VA	237.6	111	OH	0.0031	42	MI	51	43	MI	0.959
Mean			474.0			0.0032			53.1			0.983
SD			1046.9			0.000			3.5			0.029

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Table 3.6. Absolute errors for a network model based on all nodes and removal of nodes identified as important in the network based on centrality measures used to study the spread of cucurbit downy mildew in the eastern United States.

		Error after important nodes are removed based on centrality measure[b]			
Year[a]	All nodes	Betweenness	Closeness	Degree	Eigenvector
2008	0.231	0.367	0.283	0.269	0.271
2009	0.301	0.427	0.326	0.323	0.366
2010	0.353	0.443	0.365	0.365	0.365
2011	0.306	0.402	0.325	0.317	0.317
2012	0.254	0.349	0.296	0.352	0.294
2013	0.215	0.448	0.234	0.253	0.255
2014	0.168	0.307	0.198	0.247	0.247
2015	0.121	0.125	0.110	0.129	0.104
2016	0.099	0.174	0.095	0.096	0.099
Mean	**0.228**	**0.338**	**0.248**	**0.261**	**0.256**

[a] In each year, values are means of absolute errors generated across monthly time steps from January to August.

[b] A total of 20 most important nodes identified by each centrality measure were removed in the network and the model rerun to calculate the corresponding absolute errors.

Figures

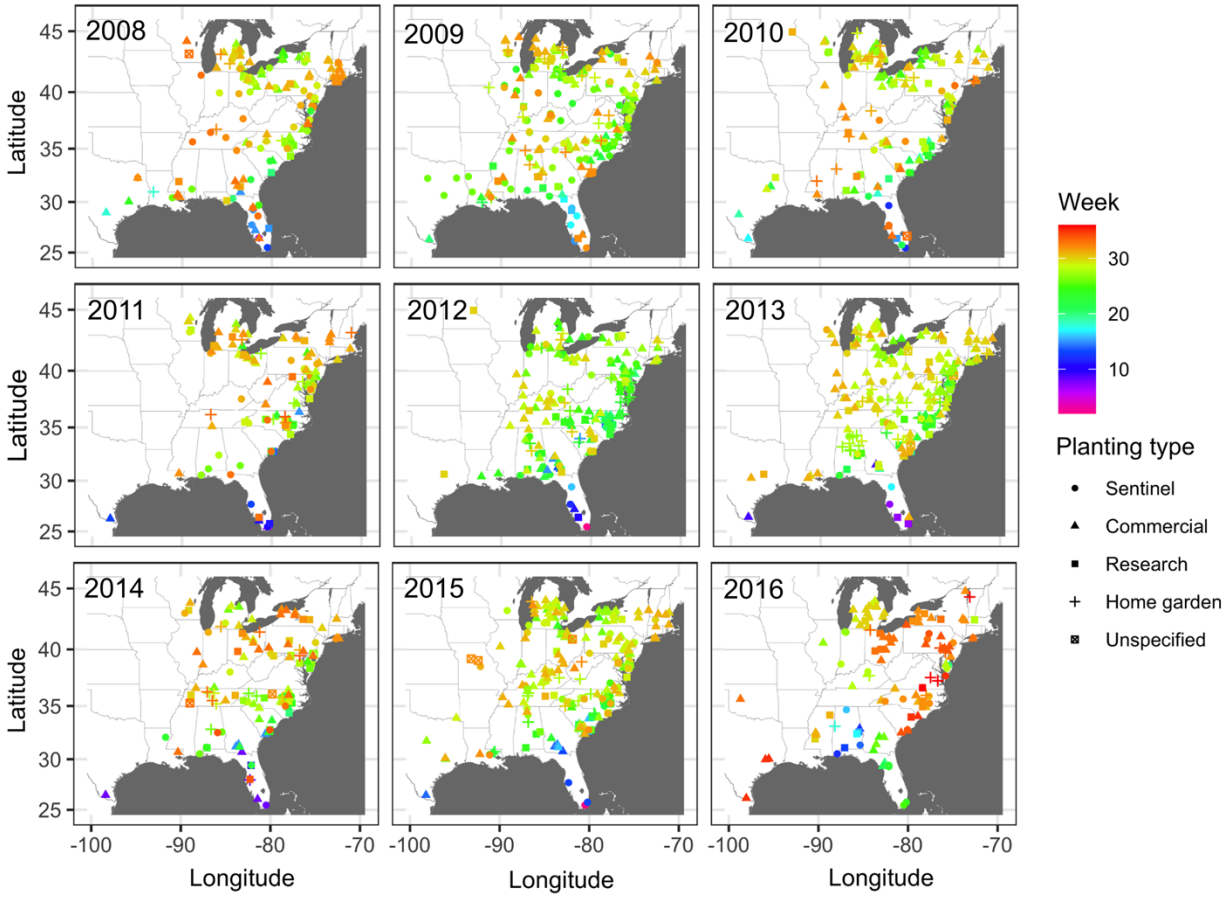


Figure 3.1. Locations of cucurbit downy mildew outbreaks in the eastern United States from 2008 to 2016. The locations are color-coded based on the week of the year. The shapes represent the planting type associated with the disease reported during the study period.

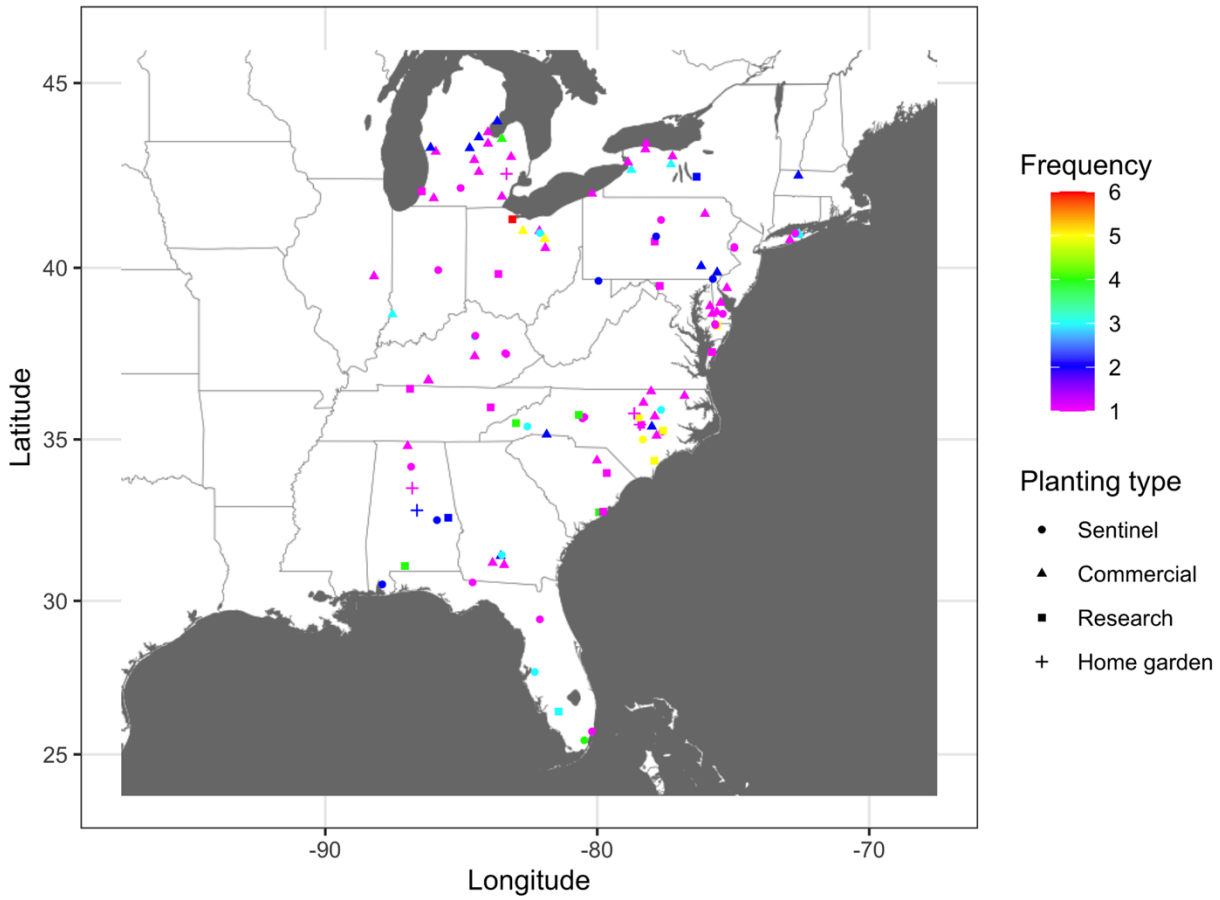


Figure 3.2. The frequency of cucurbit downy mildew outbreaks across epidemic years 2008 to 2016 in the eastern United States. Colors represent the frequency (*n*) of disease cases: red (*n* = 6), yellow (*n* = 5), green (*n* = 4), light blue (*n* = 3), blue (*n* = 2) and pink (*n* = 1). Frequency represents the number of years a node was observed as an infected node (i.e., a location where the disease was reported).

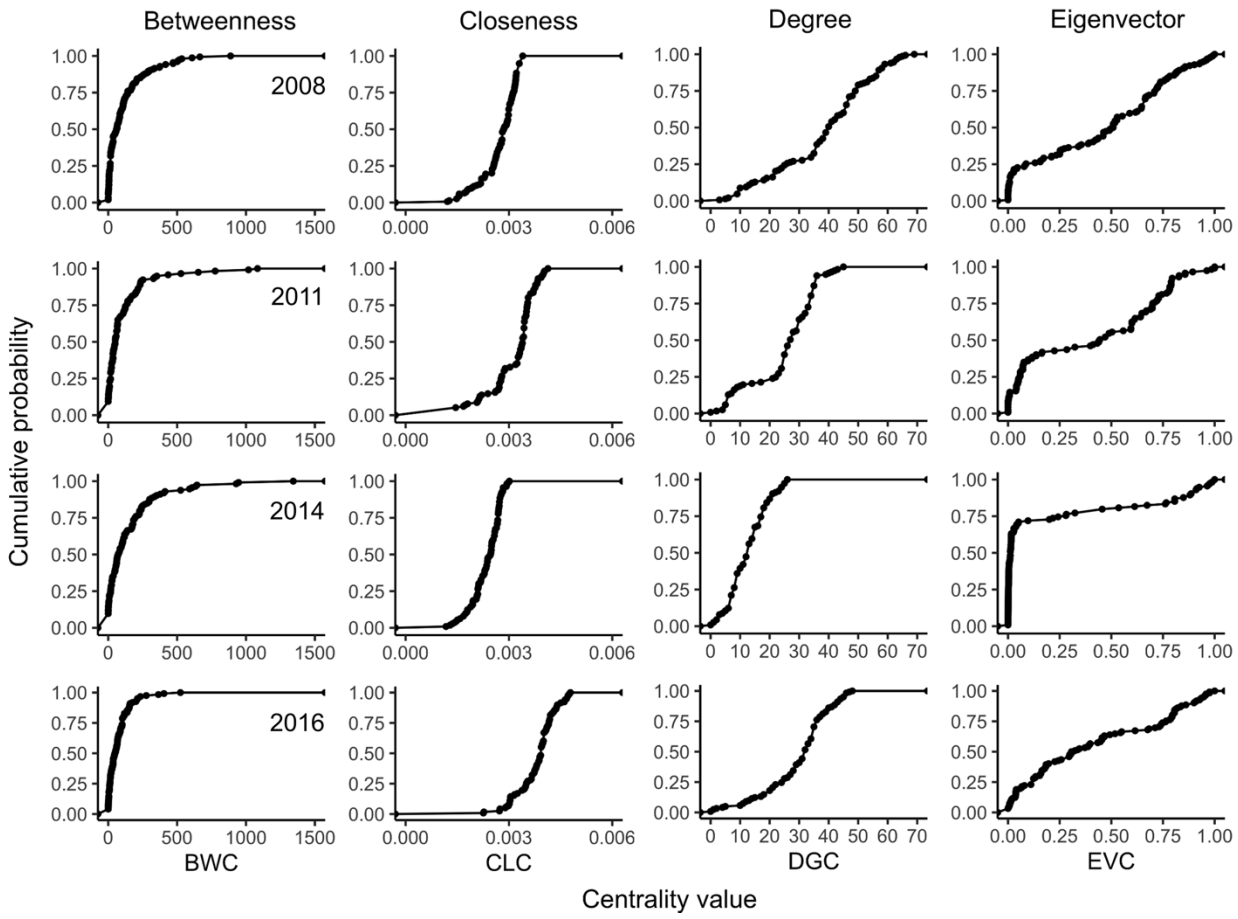


Figure 3.3. The cumulative probability distributions of centrality values of cucurbit downy mildew networks: based on disease data recorded in 2008 (first row), 2011 (second row), 2014 (third row), and 2016 (fourth row) in the eastern United States. Centrality metrics on the horizontal axis are as follows: BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality and EVC = eigenvector centrality.

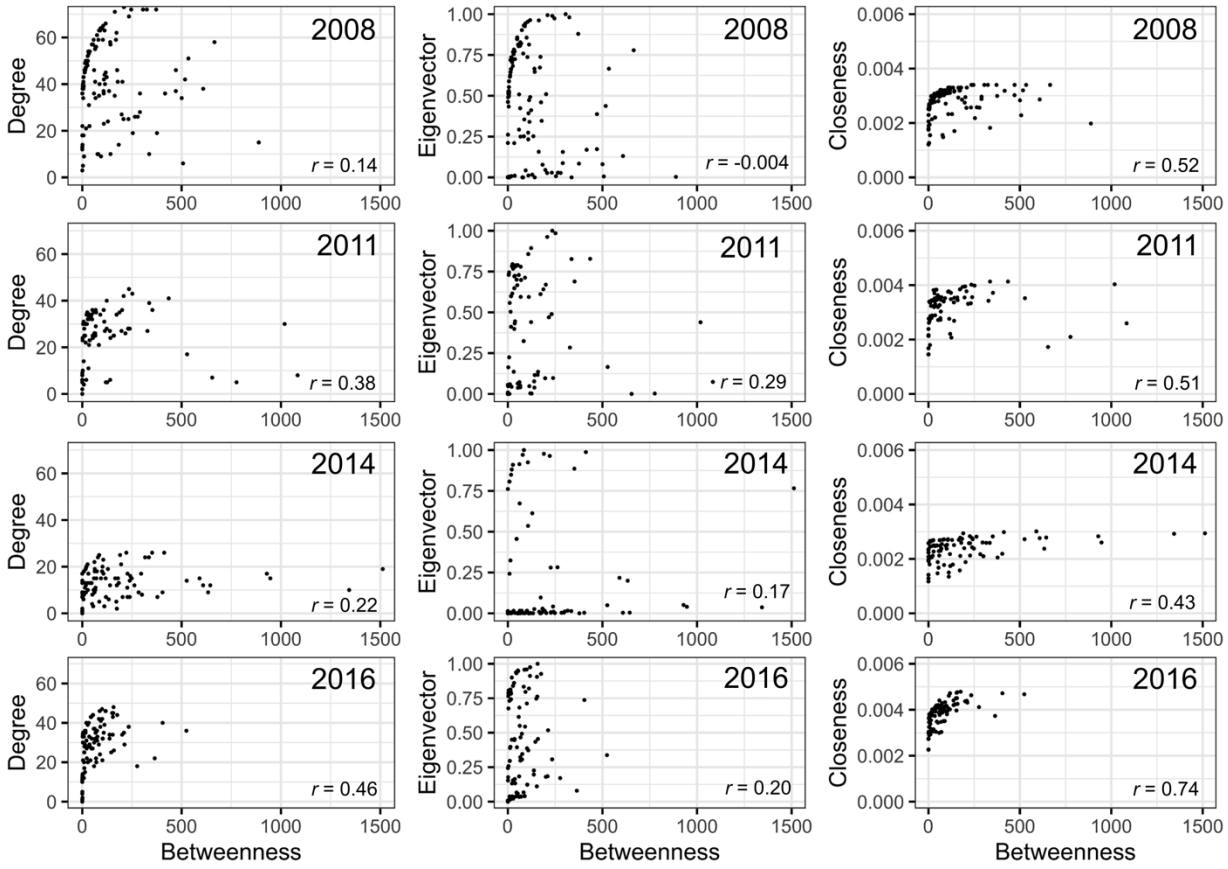


Figure 3.4. Correlation between betweenness centrality (BWC), closeness (CLC), degree (DGC), and eigenvector (EVC) centrality measures for cucurbit downy mildew networks constructed using disease data recorded in specific epidemic years in the eastern United States.

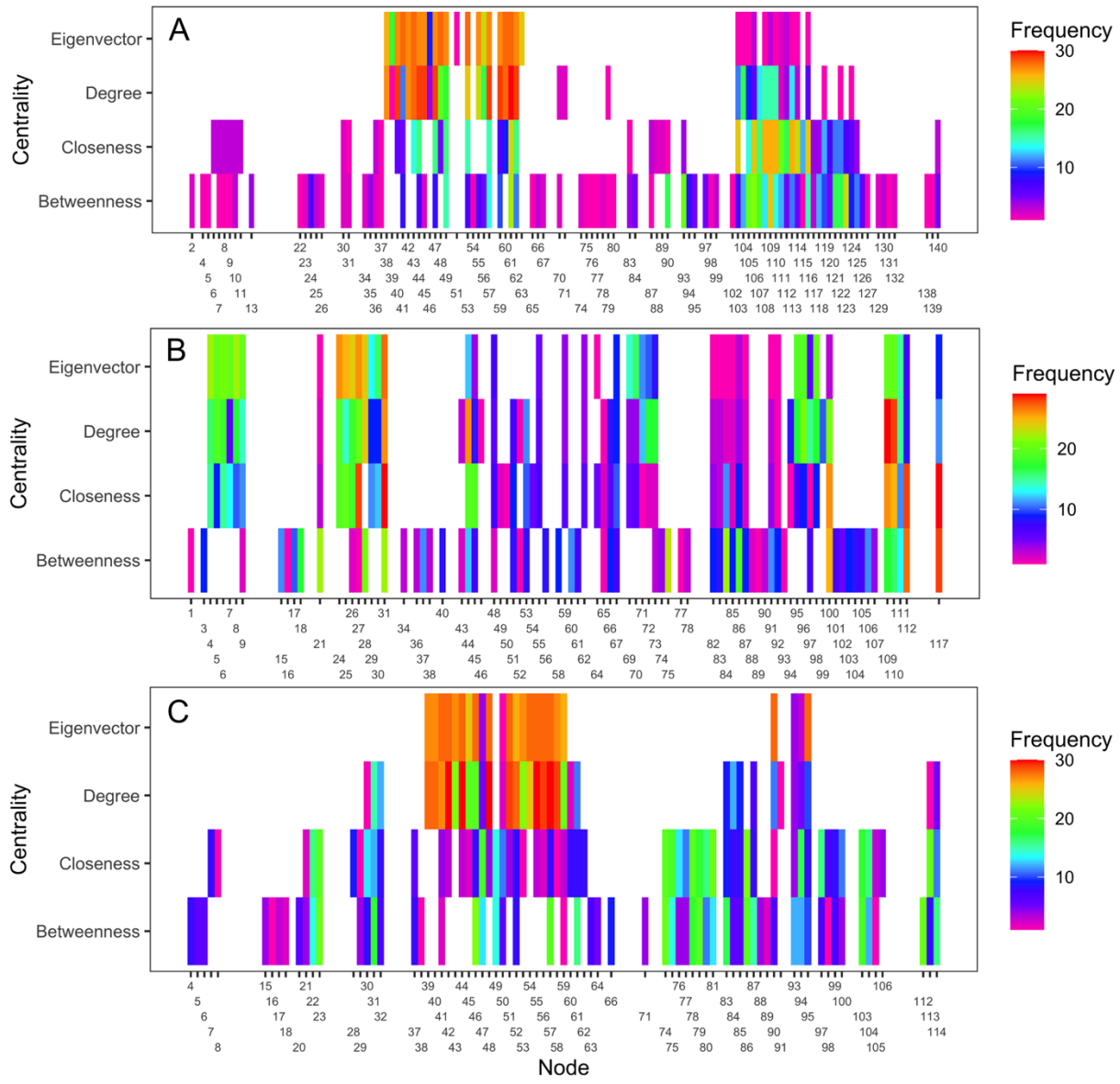


Figure 3.5. A representation of the most important nodes across 20 thresholds and the four centrality measures for 2010 (A), 2011 (B), and 2014 (C) networks. The frequency value represents the number of times a node appeared in the top 20 list across all thresholds. Most nodes overlapped across the four centrality measures in 2011. For example, node 117 in Lewis county in West Virginia appeared more than 20 times in the top 20 ranks based on BWC and CLC. This same node also appeared more than ten times in the top 20 ranks based on DGC and EVC.

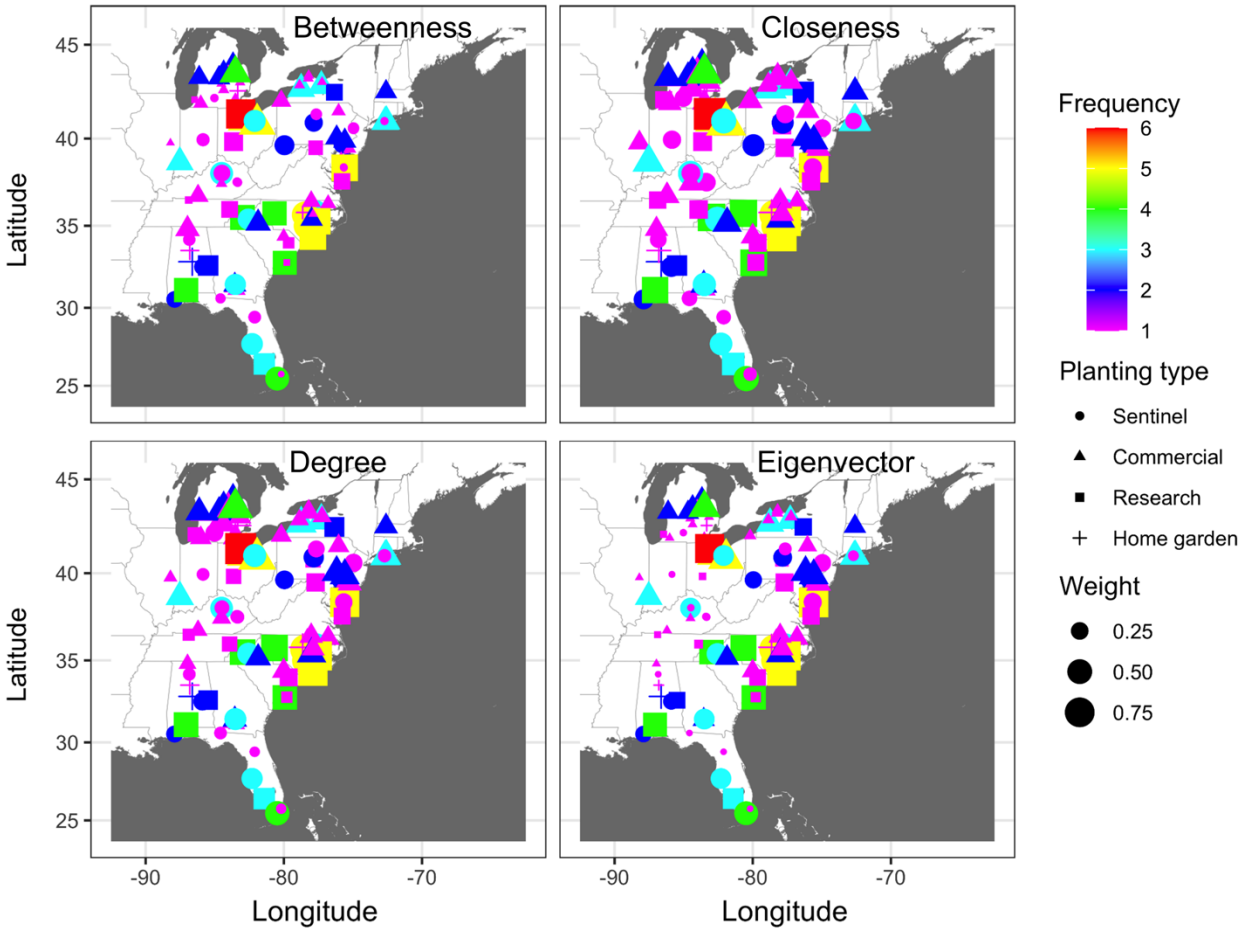


Figure 3.6. A depiction of node importance based on a combination of frequency of cucurbit downy mildew occurrence in the eastern United States and betweenness, closeness, degree, and eigenvector network centrality measures. Frequency represents the number of years a node was observed as an infected node based on epidemic years from 2008 to 2016. Frequency of occurrence and centrality measures are weighted in a ratio of 80:20.

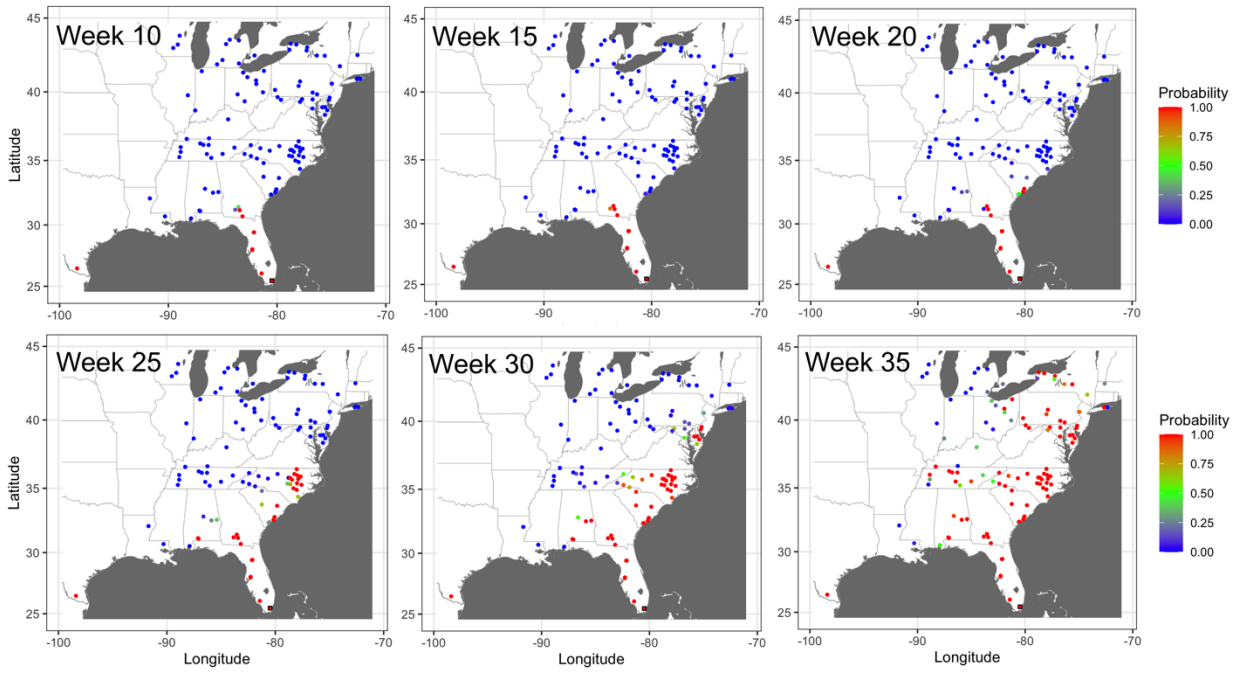


Figure 3.7. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2014 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak (The single node in Texas was reported as infected by Week 10 in the observed data; thus it considered infected with probability one by Week 10.)

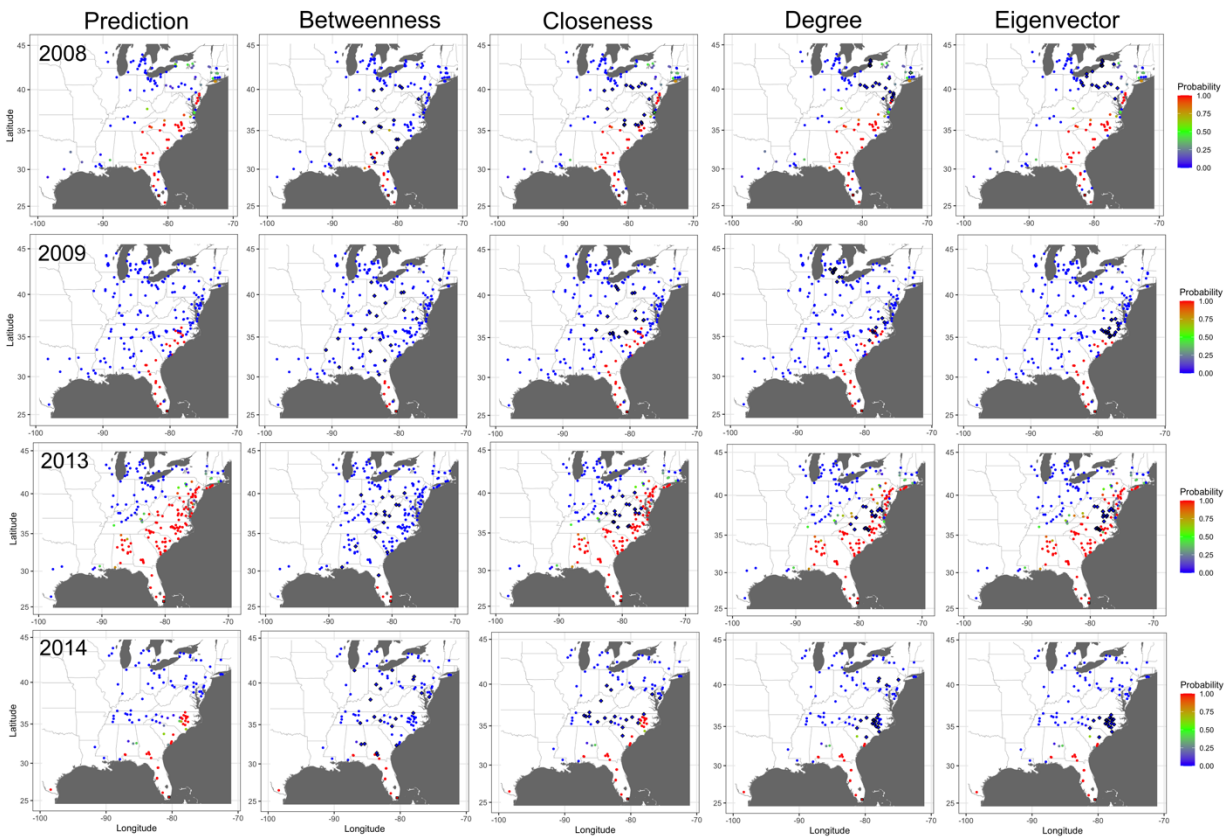


Figure 3.8. Prediction of cucurbit downy mildew outbreaks in the eastern United States by week 25 for all nodes present in the network (i.e., prediction) compared to prediction when the 20 most important nodes (based on betweenness, closeness, degree, and eigenvector centrality measures) are removed from the network. Diamond symbols are nodes identified as important based on each centrality metric.

Supplemental Tables

Table S3.1. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2011.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	18	GA	1083.8	31	MD	0.0041	109	VA	45	109	VA	1.000
2	117	WV	1018.4	67	NC	0.0041	110	VA	45	110	VA	1.000
3	17	FL	776.4	117	WV	0.0040	28	MD	43	28	MD	0.985
4	15	FL	654.0	109	VA	0.0040	111	VA	42	111	VA	0.962
5	21	KY	527.6	110	VA	0.0040	67	NC	41	44	NC	0.894
6	67	NC	435.0	28	MD	0.0040	44	NC	40	30	MD	0.857
7	100	PA	352.7	45	NC	0.0040	31	MD	39	29	MD	0.857
8	31	MD	336.7	53	NC	0.0039	29	MD	36	67	NC	0.828
9	75	NY	328.1	51	NC	0.0038	30	MD	36	31	MD	0.827
10	28	MD	251.8	66	NC	0.0038	45	NC	36	8	DE	0.795
11	82	OH	239.0	24	MD	0.0038	71	NJ	36	5	DE	0.795
12	109	VA	235.0	27	MD	0.0038	95	PA	36	9	DE	0.795
13	110	VA	235.0	26	MD	0.0038	96	PA	36	95	PA	0.790
14	103	SC	231.3	111	VA	0.0038	98	PA	36	98	PA	0.790
15	104	SC	231.3	112	VA	0.0038	100	PA	36	96	PA	0.790
16	105	SC	231.3	29	MD	0.0038	5	DE	35	24	MD	0.789
17	101	SC	216.7	30	MD	0.0038	8	DE	35	69	NJ	0.783
18	102	SC	216.7	44	NC	0.0037	9	DE	35	70	NJ	0.783
19	111	VA	208.1	100	PA	0.0037	53	NC	35	27	MD	0.782
20	45	NC	200.6	25	MD	0.0037	55	NC	35	71	NJ	0.780
Mean			400.5			0.0039			37.9			0.844
SD			273.1			0.0000			3.48			0.080

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Table S3.2. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2012.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	164	TN	3587.1	82	NC	0.0015	158	SC	33	81	NC	1.000
2	168	VA	2860.4	101	NC	0.0015	81	NC	30	90	NC	0.985
3	8	AL	2421.8	85	NC	0.0015	160	SC	30	87	NC	0.985
4	101	NC	1652.5	102	NC	0.0015	79	NC	29	79	NC	0.985
5	38	IN	1635.0	91	NC	0.0015	82	NC	29	82	NC	0.977
6	11	AL	1622.0	110	NC	0.0015	87	NC	29	86	NC	0.954
7	39	IN	1602.9	81	NC	0.0015	90	NC	29	109	NC	0.920
8	171	VA	1528.9	94	NC	0.0015	16	DE	28	85	NC	0.918
9	102	NC	1515.5	79	NC	0.0015	65	MD	28	98	NC	0.915
10	48	KY	1500.0	87	NC	0.0015	86	NC	28	84	NC	0.915
11	82	NC	1435.0	90	NC	0.0015	109	NC	28	89	NC	0.915
12	173	WV	1408.8	99	NC	0.0015	64	MD	27	97	NC	0.915
13	107	NC	1252.5	168	VA	0.0015	66	MD	27	110	NC	0.912
14	104	NC	1159.8	86	NC	0.0015	85	NC	27	91	NC	0.912
15	36	GA	1127.1	104	NC	0.0015	91	NC	27	95	NC	0.891
16	106	NC	1086.2	107	NC	0.0015	95	NC	27	92	NC	0.871
17	85	NC	1043.3	108	NC	0.0015	101	NC	27	83	NC	0.871
18	94	NC	1013.2	106	NC	0.0014	110	NC	27	80	NC	0.871
19	165	TN	970.0	158	SC	0.0014	137	PA	27	93	NC	0.848
20	6	AL	920.6	96	NC	0.0014	139	PA	27	88	NC	0.831
Mean			1567.1			0.0015			28.9			0.919
SD			672.6			0.0000			1.54			0.049

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Table S3.3. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2013.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	203	WV	863.5	203	WV	0.0026	203	WV	70	189	VA	1.000
2	25	FL	784.0	191	VA	0.0025	196	VA	68	192	VA	1.000
3	202	WV	755.6	202	WV	0.0025	189	VA	64	196	VA	0.991
4	191	VA	741.0	196	VA	0.0025	191	VA	64	203	WV	0.979
5	165	SC	693.7	201	WV	0.0025	192	VA	64	190	VA	0.974
6	204	WV	644.6	106	NC	0.0025	87	NC	62	186	VA	0.968
7	196	VA	538.9	102	NC	0.0024	98	NC	62	198	VA	0.964
8	36	IN	501.3	189	VA	0.0024	99	NC	61	188	VA	0.960
9	158	PA	490.5	192	VA	0.0024	188	VA	61	193	VA	0.943
10	176	SC	446.2	39	KY	0.0024	190	VA	61	65	MD	0.926
11	138	OH	425.9	204	WV	0.0024	106	NC	60	63	MD	0.926
12	39	KY	419.1	99	NC	0.0024	186	VA	60	87	NC	0.922
13	109	NC	415.5	190	VA	0.0024	198	VA	60	98	NC	0.922
14	10	AL	411.5	40	KY	0.0024	63	MD	59	194	VA	0.912
15	2	AL	404.1	188	VA	0.0024	65	MD	59	89	NC	0.900
16	106	NC	401.1	158	PA	0.0024	85	NC	59	92	NC	0.900
17	201	WV	392.2	198	VA	0.0024	86	NC	59	85	NC	0.900
18	104	NC	376.9	87	NC	0.0024	89	NC	59	86	NC	0.900
19	14	AL	367.4	98	NC	0.0024	91	NC	59	107	NC	0.900
20	156	PA	355.5	138	OH	0.0024	92	NC	59	91	NC	0.900
Mean			521.4			0.0002			61.5			0.939
SD			162.4			0.0000			3.12			0.037

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Table S3.4. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2014.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	52	NC	1512.3	103	TN	0.0030	46	NC	26	51	NC	1.000
2	23	KY	1343.3	46	NC	0.0030	90	SC	26	56	NC	1.000
3	99	TN	946.4	52	NC	0.0029	95	SC	26	46	NC	0.988
4	30	MD	928.9	57	NC	0.0029	51	NC	25	57	NC	0.977
5	74	OH	644.1	23	KY	0.0029	56	NC	25	58	NC	0.971
6	88	SC	633.3	30	MD	0.0028	57	NC	25	90	SC	0.964
7	79	OH	608.2	97	TN	0.0028	58	NC	24	39	NC	0.925
8	103	TN	590.1	95	SC	0.0028	83	PA	24	40	NC	0.914
9	29	MD	525.6	49	NC	0.0028	87	PA	24	48	NC	0.910
10	46	NC	412.4	74	OH	0.0028	39	NC	23	55	NC	0.910
11	4	AL	403.7	104	AL	0.0028	40	AL	22	44	AL	0.910
12	20	IN	378.0	79	OH	0.0028	44	NC	21	95	SC	0.886
13	95	SC	351.6	100	TN	0.0028	48	NC	21	41	NC	0.881
14	83	PA	335.1	22	KY	0.0027	55	NC	21	42	NC	0.881
15	87	PA	316.1	98	TN	0.0027	61	NJ	21	54	NC	0.849
16	17	GA	300.4	106	TN	0.0027	41	NC	20	59	NC	0.807
17	18	GA	300.4	29	MD	0.0027	42	NC	20	43	NC	0.807
18	80	OH	296.0	93	SC	0.0027	84	PA	20	53	NC	0.807
19	6	AL	283.5	112	AL	0.0027	52	AL	19	52	AL	0.766
20	49	NC	262.4	51	NC	0.0027	54	NC	19	45	NC	0.761
Mean			568.6			0.0003			22.6			0.896
SD			357.1			0.0000			2.458			0.076

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Table S3.5. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2015.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	1	AL	1266.4	208	WV	0.0027	207	WV	90	207	WV	1.000
2	208	WV	911.2	207	WV	0.0027	208	WV	87	208	WV	0.952
3	3	AL	743.9	206	WV	0.0026	99	NC	82	55	MD	0.930
4	207	WV	730.4	99	NC	0.0026	182	SC	81	58	MD	0.864
5	206	WV	712.2	93	NC	0.0026	175	SC	80	163	PA	0.852
6	99	NC	595.7	142	OH	0.0026	206	WV	79	57	MD	0.844
7	93	NC	575.8	209	WV	0.0026	93	NC	78	159	PA	0.840
8	175	SC	520.3	175	SC	0.0025	176	SC	78	153	PA	0.820
9	182	SC	477.5	138	OH	0.0025	55	MD	77	12	DE	0.811
10	142	OH	471.1	143	OH	0.0025	149	PA	77	53	MD	0.810
11	138	AL	460.6	182	SC	0.0025	94	NC	76	151	PA	0.808
12	143	OH	448.9	38	KY	0.0025	96	NC	73	52	MD	0.807
13	17	FL	412.0	176	SC	0.0025	163	PA	73	96	NC	0.807
14	149	PA	390.3	29	KY	0.0025	89	NC	72	155	PA	0.804
15	47	LA	390.2	36	KY	0.0025	158	PA	72	201	VA	0.798
16	49	LA	390.2	94	NC	0.0025	159	PA	72	152	PA	0.797
17	176	SC	382.6	141	OH	0.0025	100	NC	71	160	PA	0.797
18	96	NC	375.3	37	KY	0.0025	134	OH	71	13	DE	0.794
19	45	AL	358.0	39	KY	0.0025	144	OH	71	148	PA	0.792
20	24	IL	346.5	40	KY	0.0025	58	MD	70	14	DE	0.787
Mean			548.0			0.0003			76.5			0.836
SD			230.1			0.0000			5.549			0.059

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Table S3.6. Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for epidemic data observed in the eastern United States in 2016.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	53	NC	523.9	86	OH	0.0048	90	OH	48	90	OH	1.000
2	88	OH	404.5	90	OH	0.0048	70	NY	47	92	OH	0.975
3	6	AL	364.5	82	OH	0.0047	73	NY	46	84	OH	0.975
4	25	KY	276.3	84	OH	0.0047	84	OH	46	70	NY	0.959
5	118	VA	233.8	88	OH	0.0047	92	OH	46	100	PA	0.957
6	119	VA	233.8	92	OH	0.0047	100	PA	46	91	OH	0.952
7	26	KY	212.5	53	NC	0.0047	94	PA	45	85	OH	0.947
8	58	NC	209.6	118	VA	0.0046	101	PA	45	87	OH	0.937
9	59	NC	209.6	119	VA	0.0046	85	OH	44	89	OH	0.931
10	57	NC	200.6	81	OH	0.0046	86	OH	44	86	OH	0.927
11	86	OH	175.7	94	PA	0.0045	91	OH	44	73	NY	0.912
12	27	KY	158.9	101	PA	0.0045	82	OH	43	82	OH	0.906
13	90	OH	157.2	63	NC	0.0044	87	OH	43	43	MI	0.900
14	101	PA	155.6	64	NC	0.0044	89	OH	43	46	MI	0.900
15	82	OH	153.9	58	NC	0.0044	69	NY	42	83	OH	0.861
16	105	SC	153.0	59	NC	0.0044	97	PA	42	41	MI	0.840
17	94	PA	150.4	57	NC	0.0044	75	NY	41	69	NY	0.834
18	28	Ky	145.4	102	PA	0.0043	43	MI	40	81	OH	0.821
19	55	NC	137.5	26	KY	0.0043	46	MI	40	42	MI	0.813
20	62	NC	137.5	103	PA	0.0043	83	OH	40	39	MI	0.807
Mean			219.7			0.0005			43.8			0.908
SD			102.1			0.0000			2.403			0.059

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation

Supplemental Figures

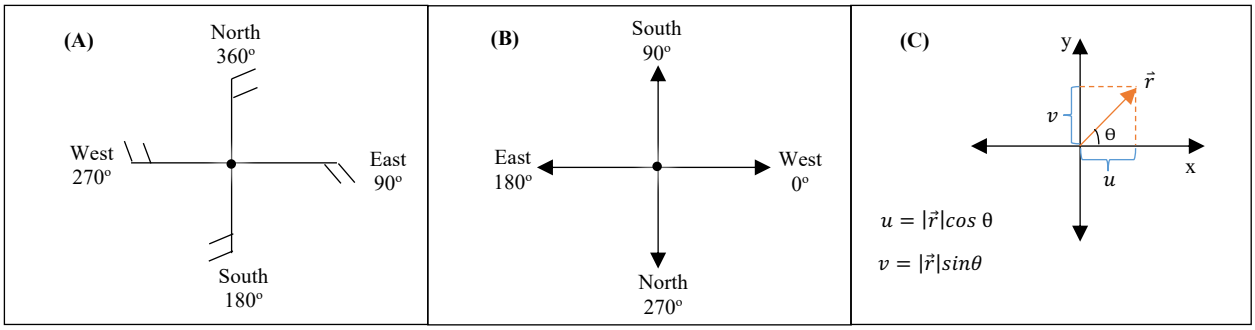


Figure S3.1. A graphical illustration of the conversion of meteorological wind direction to mathematical wind direction. (A) A wind bar graph representing raw observations for the meteorological wind direction, i.e., a north wind is 360°, a south wind is 180°, a west wind is 270°, and an east wind is 90°. **(B)** The mathematical convention for the meteorological wind direction, i.e., a north wind is 270°, a south wind is 90°, a west wind is 0°, and an east wind is 180°. **(C)** The u and v components of the wind where $|\vec{r}|$ is the magnitude (or wind speed). Here, we treat $|\vec{r}|$ as the observed wind speed in miles per hour. A positive u and a negative u represent a west wind and an east wind, respectively. A positive v and a negative v represent a south wind and a north wind, respectively. The figures and notes are courtesy of Virginia Weather & Climate Data lecture notes.

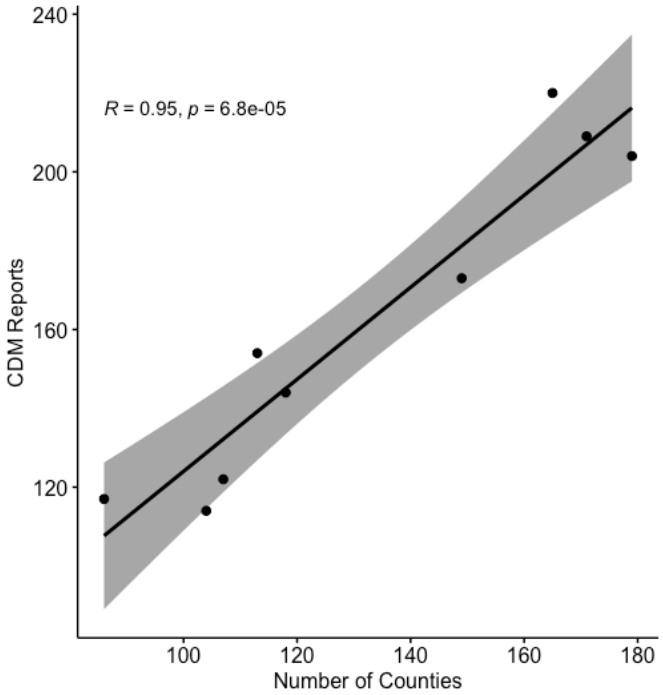


Figure S3.2. A Pearson correlation analysis of the number of counties and the total number of CDM reports recorded from 2008 to 2016.

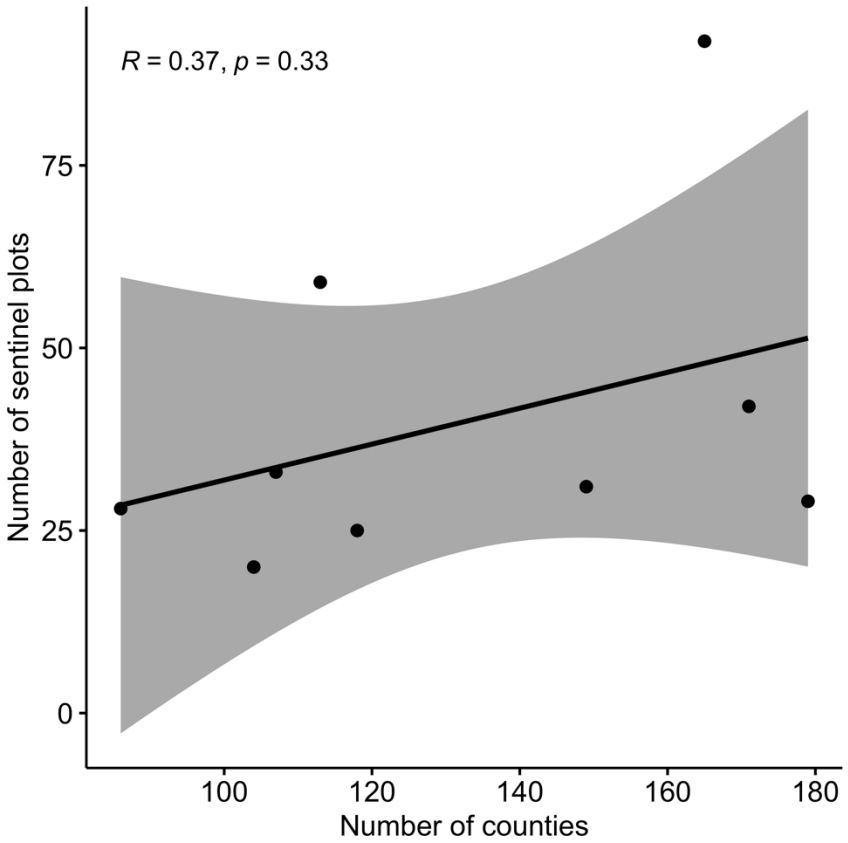


Figure S3.3. A Pearson correlation analysis of the number of counties and the locations with active surveillance in 2008 to 2016.

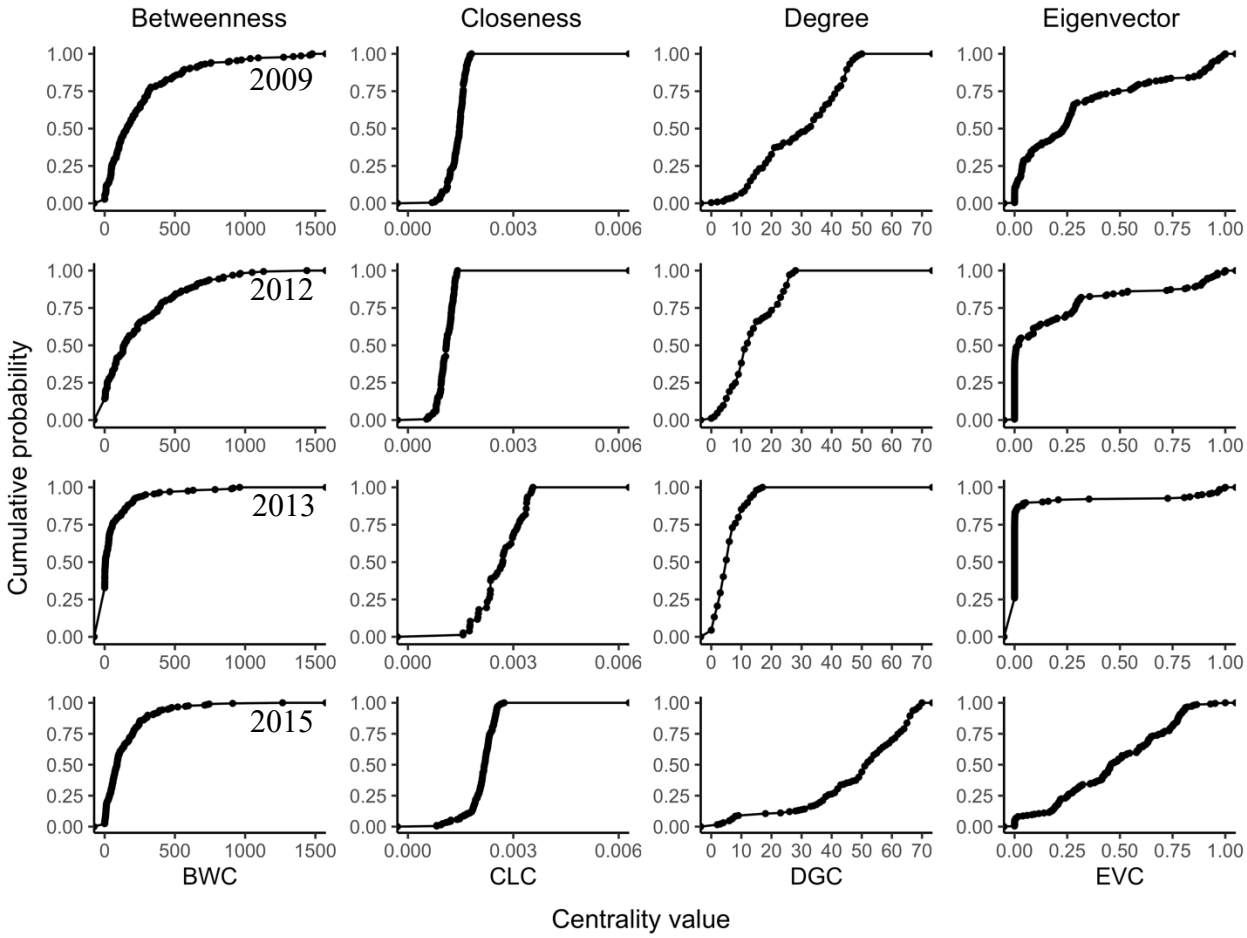


Figure S3.4. The cumulative probability distributions of centrality values of cucurbit downy mildew networks (2009, 2012, 2013, and 2015): based on disease data recorded in 2009 (first row), 2012 (second row), 2013 (third row), and 2015 (fourth row) in the eastern United States. Centrality metrics on the horizontal axis are as follows: BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality and EVC = eigenvector centrality.

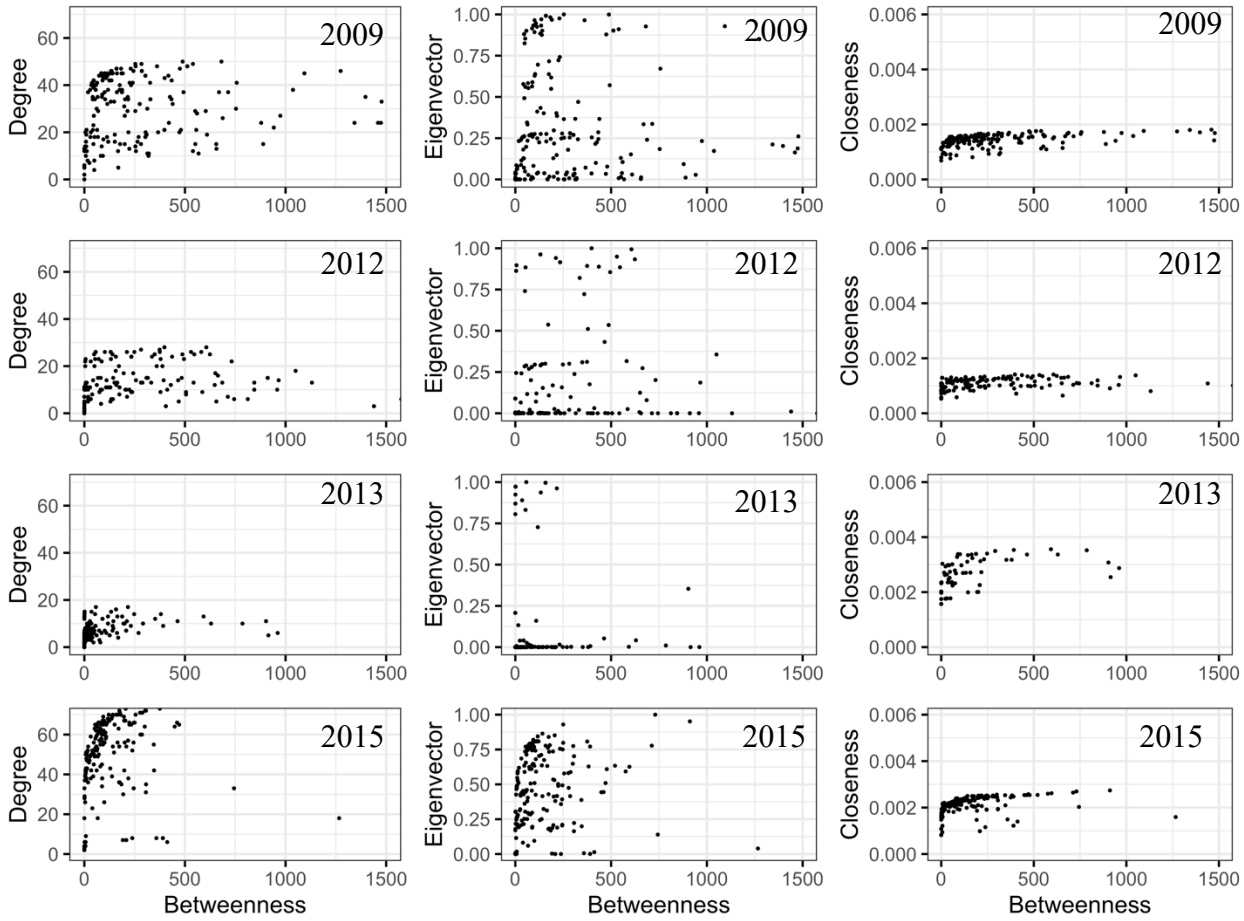


Figure S3.5. Correlation between betweenness centrality (BWC) and closeness (CLC), degree (DGC), and eigenvector (EVC) centrality measures for networks of cucurbit downy mildew constructed using disease data recorded in specific epidemic years in the eastern United States.

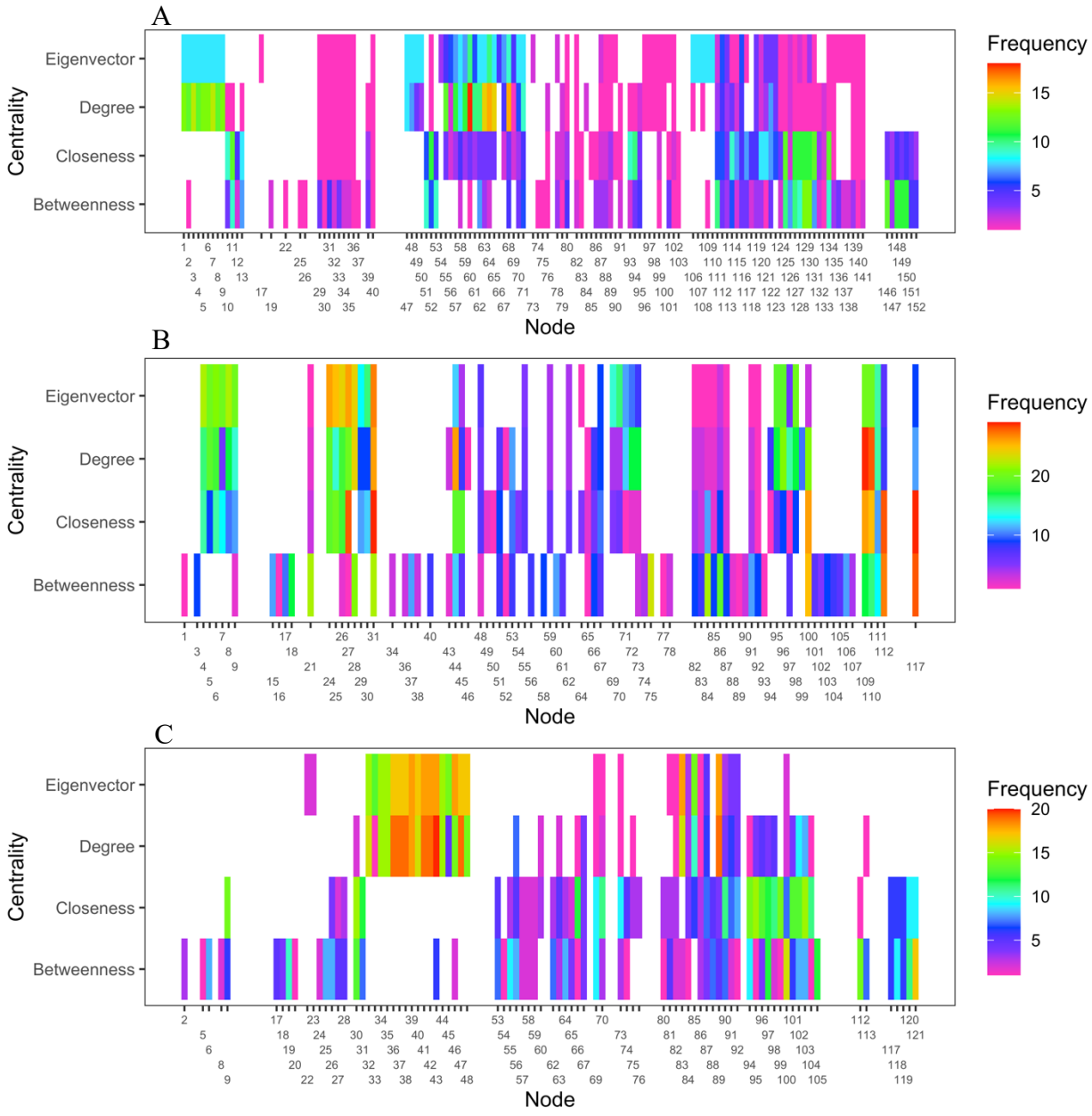


Figure S3.6. A representation of the most important nodes across 20 thresholds and the four centrality measures for 2008 (A), 2011 (B), and 2016 (C) networks. The frequency value represents the number of times a node appeared in the top 20 list across all thresholds. Most nodes overlapped across the four centrality measures in 2011. For example, node 117 in Lewis county in West Virginia appeared more than 20 times in the top 20 ranks based on BWC and CLC. This same node also appeared more than ten times in the top 20 ranks based on DGC and EVC.

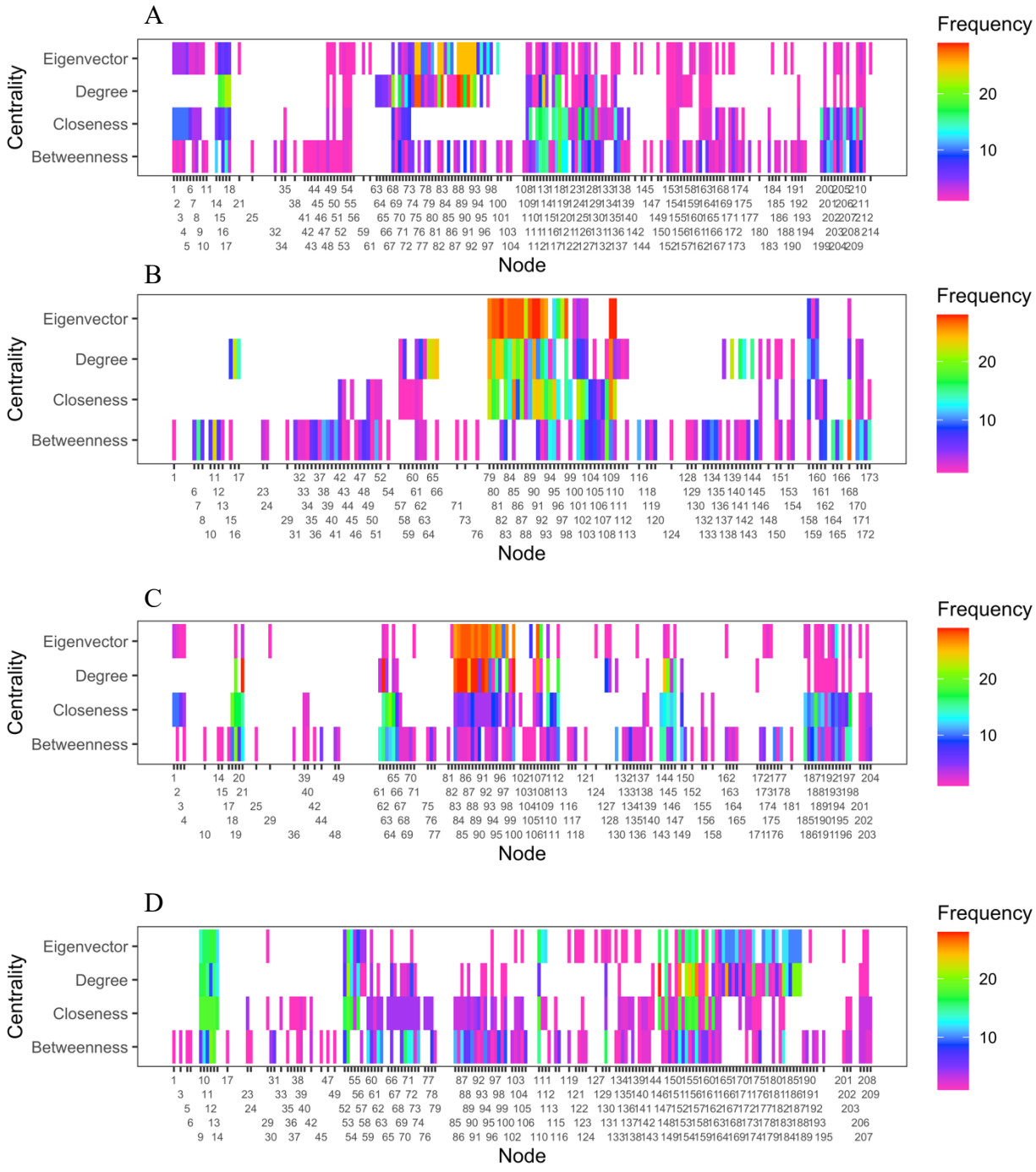


Figure S3.7. A representation of the most important nodes across 20 thresholds and the four centrality measures for 2009 (A), 2012 (B), 2013 (C), and 2015 (D) networks. The frequency value represents the number of times a node appeared in the top 20 list across all thresholds. Most nodes overlapped across the four centrality measures in 2011. For example, node 117 in Lewis county in West Virginia appeared more than 20 times in the top 20 ranks based on BWC and CLC. This same node also appeared more than ten times in the top 20 ranks based on DGC and EVC.

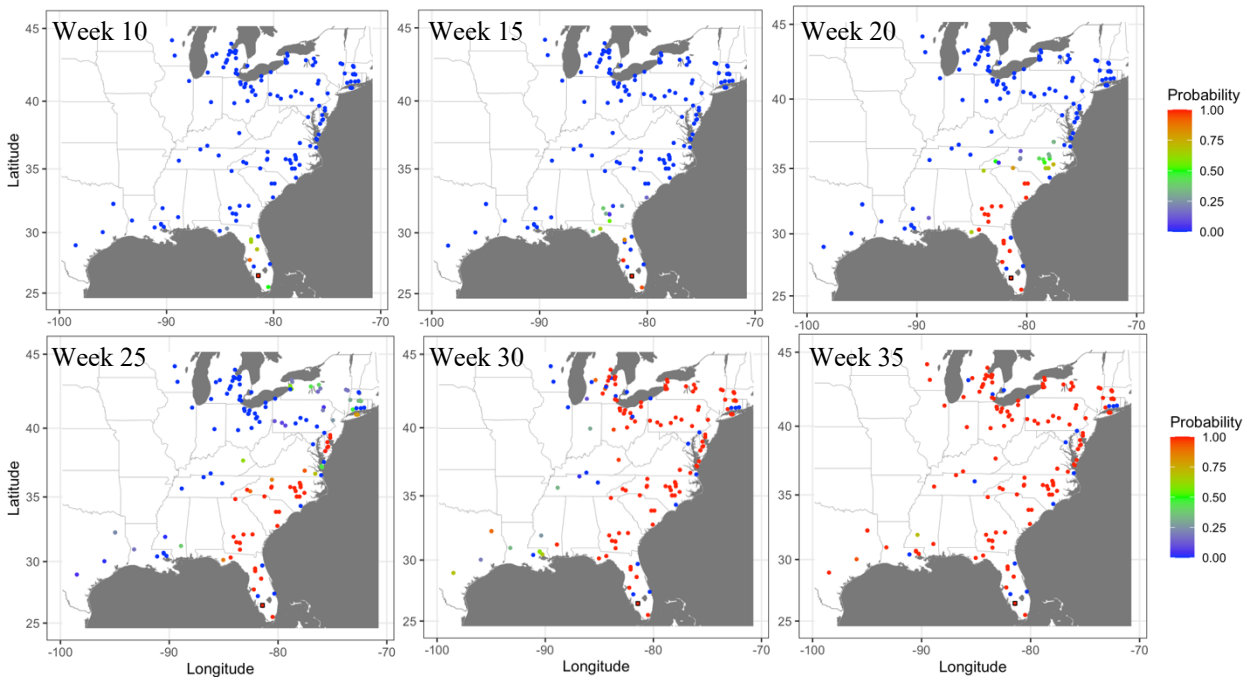


Figure S3.8. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2008 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

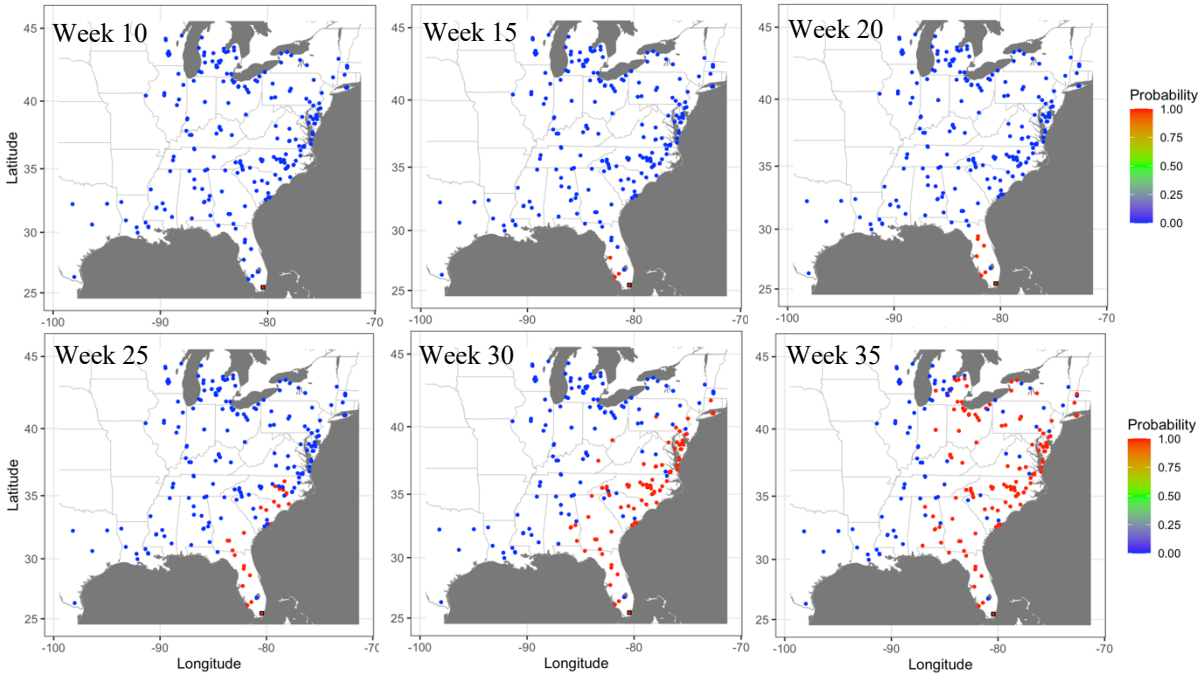


Figure S3.9. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2009 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

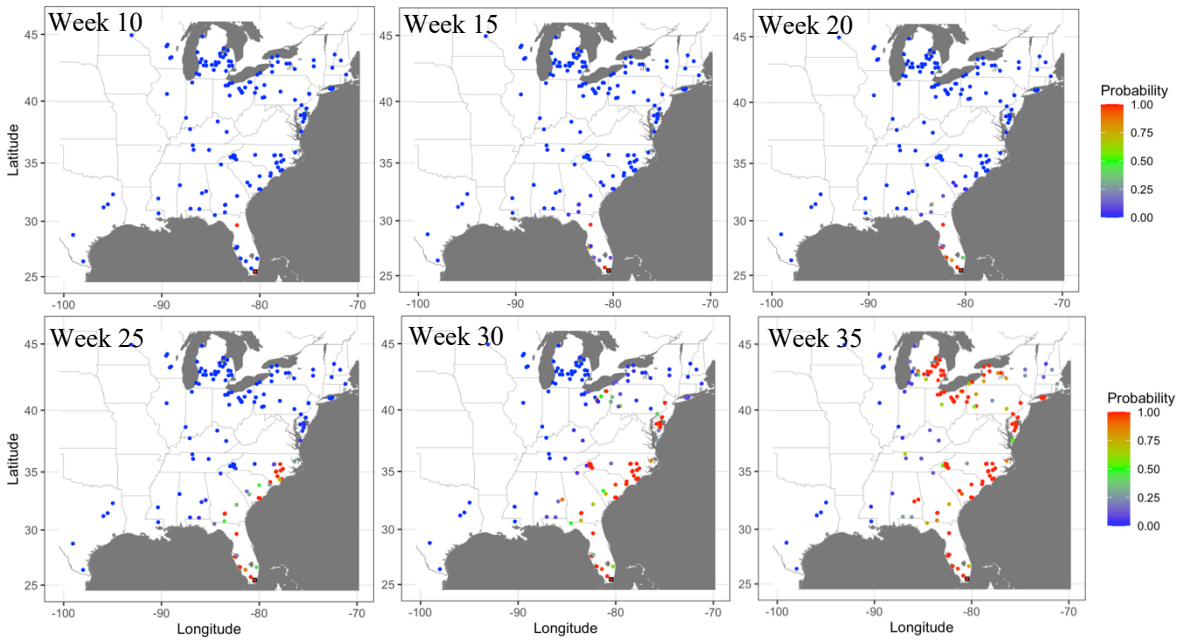


Figure S3.10. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2010 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

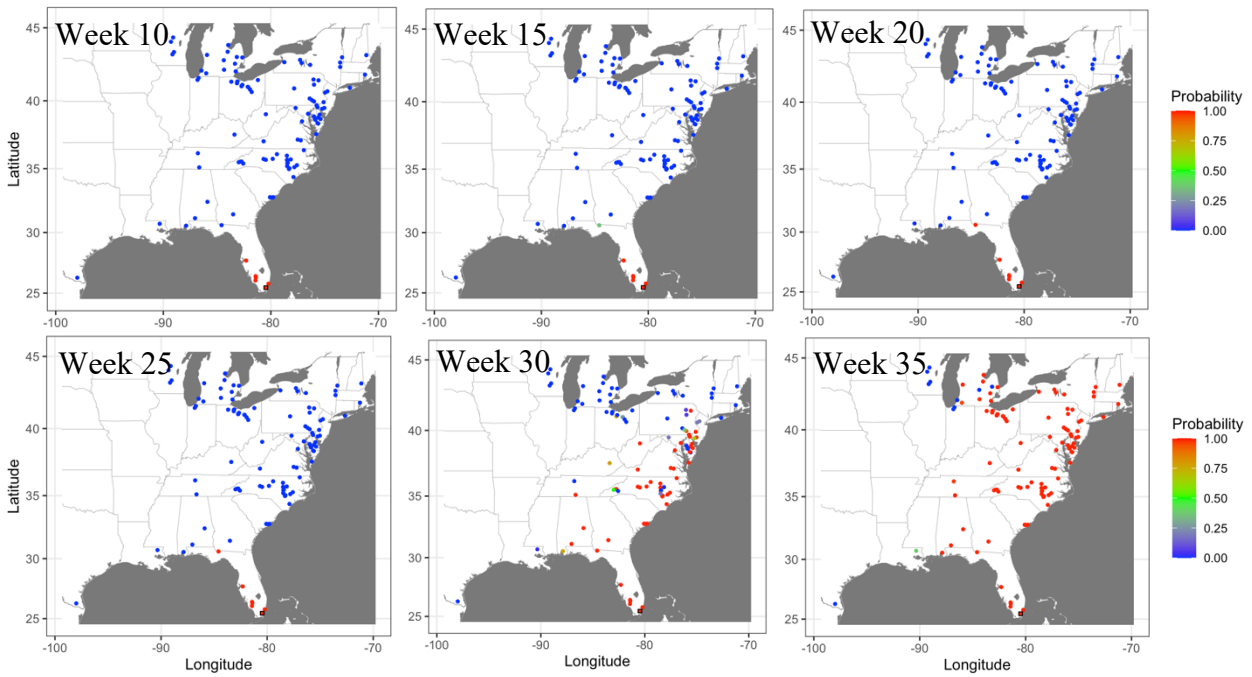


Figure S3.11. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2011 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

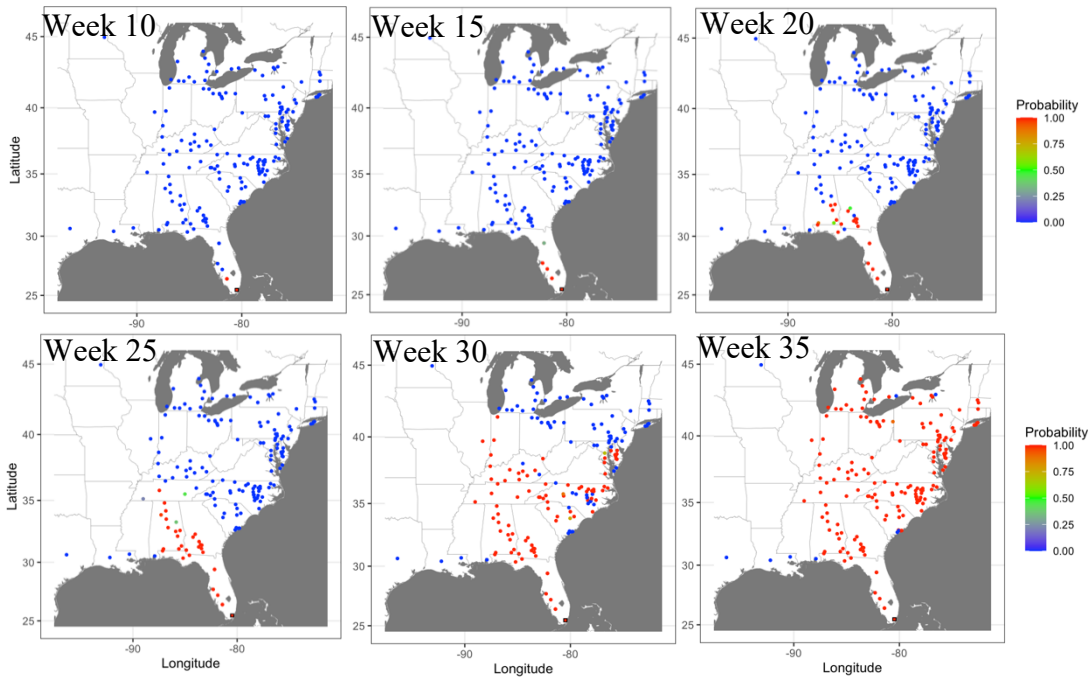


Figure S3.12. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2012 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

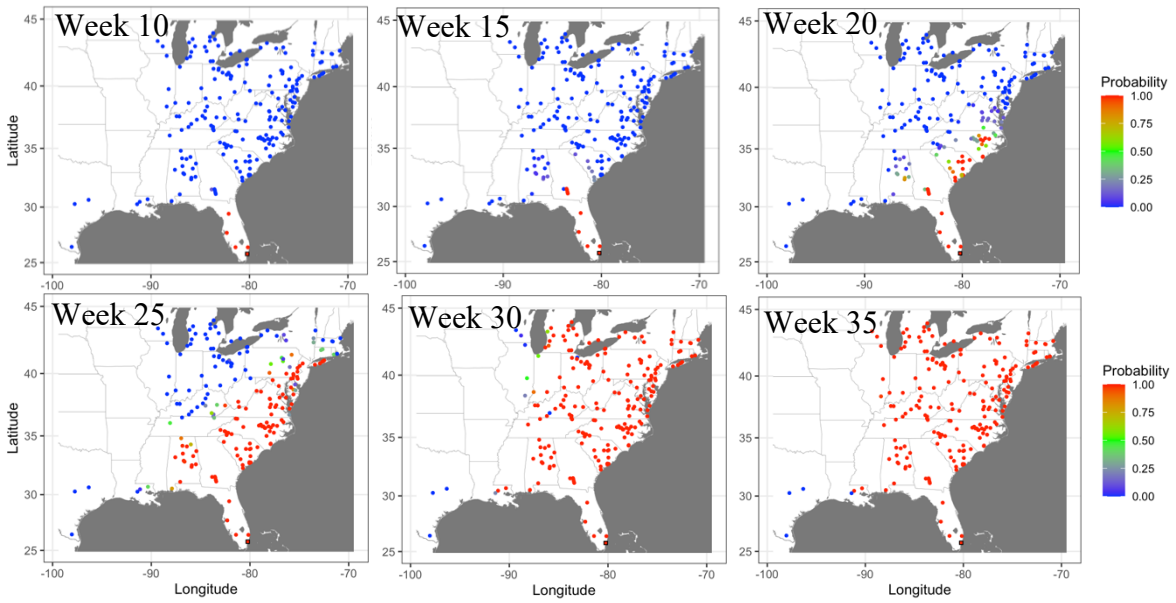


Figure S3.13. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2013 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

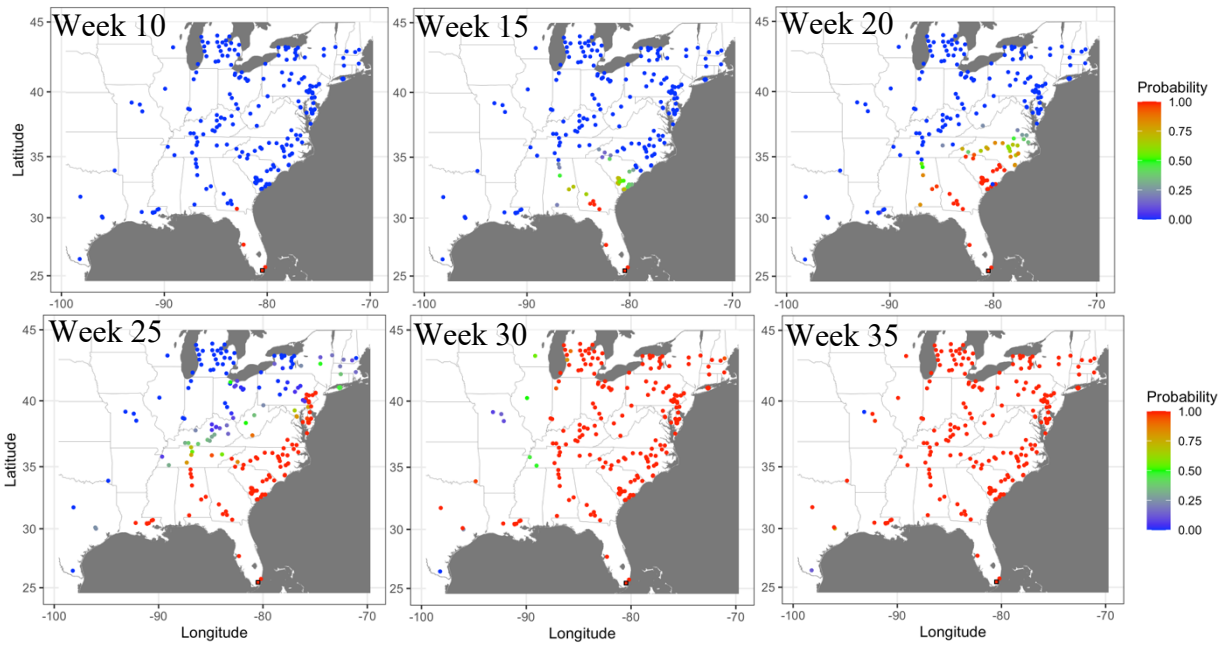


Figure S3.14. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2015 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

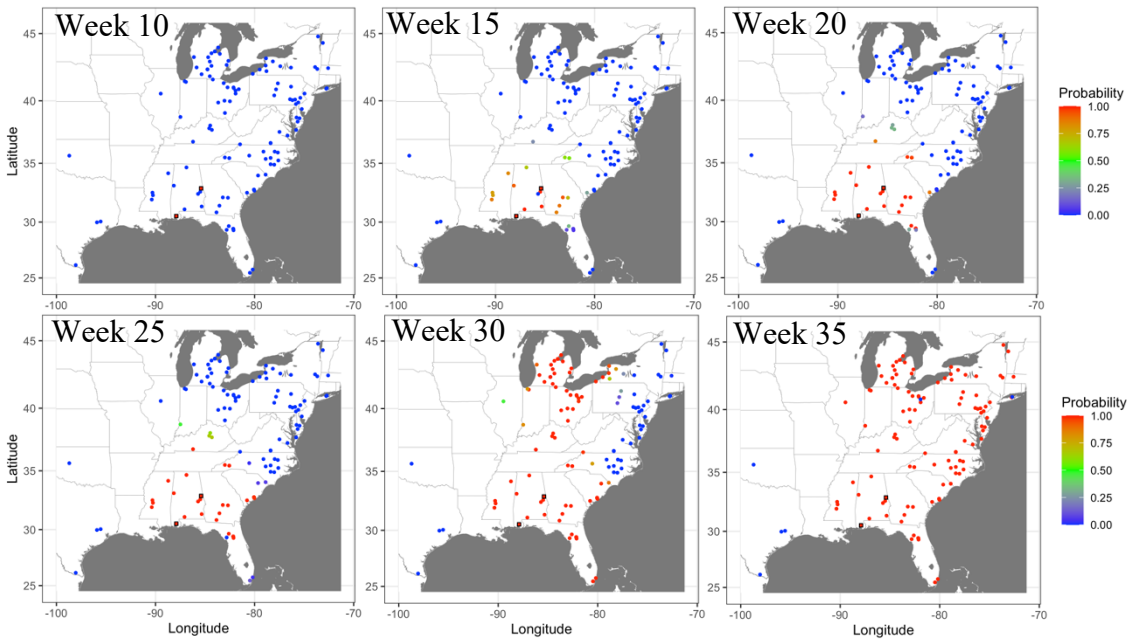


Figure S3.15. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2016 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with high probability. Blue nodes represent counties predicted to be no outbreak with negligible probability of infection, and all other shades from green to dark red represent the increasing probability of disease outbreak.

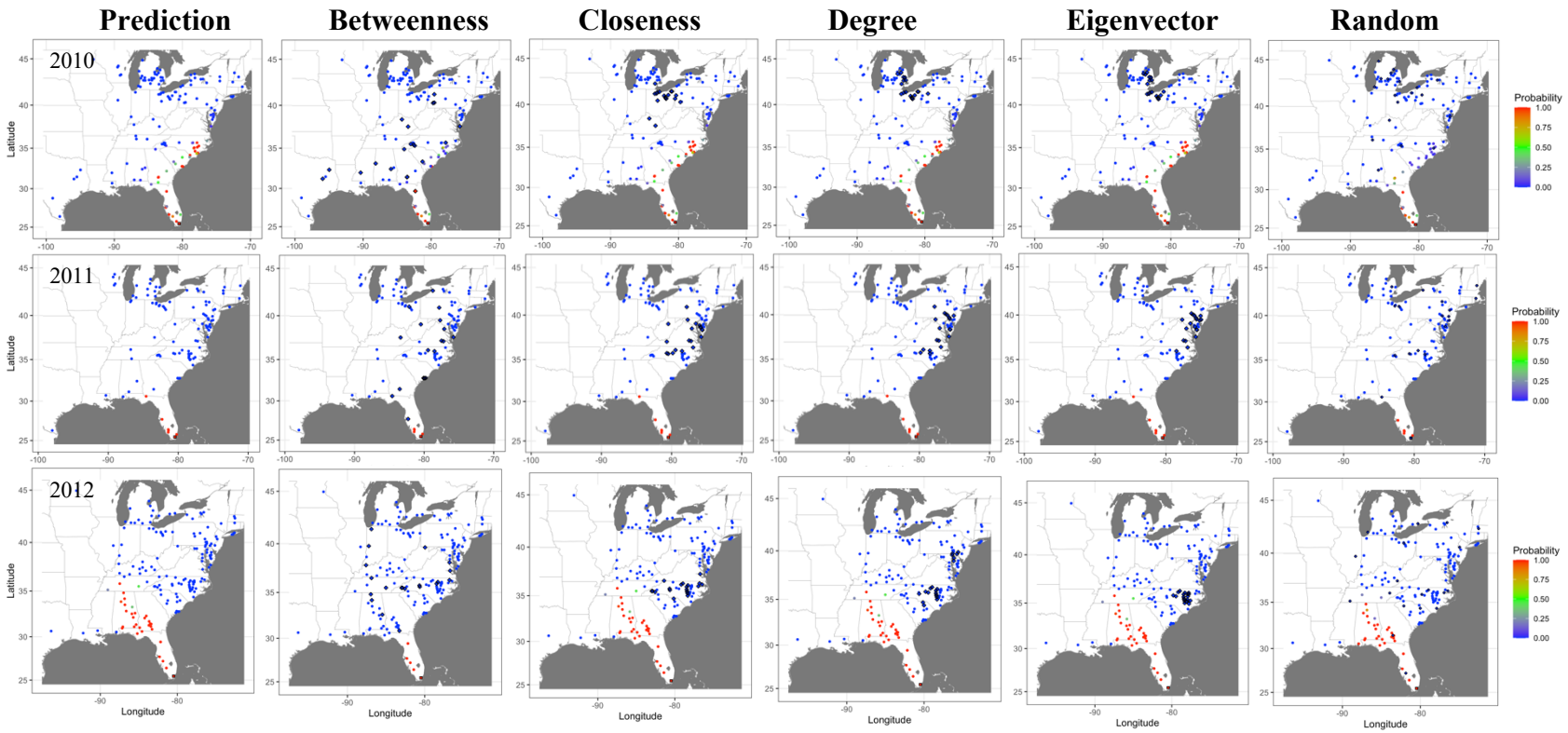


Figure S3.16. Prediction of cucurbit downy mildew outbreaks in the eastern United States by week 25 for all nodes present in the network (i.e., prediction) compared to prediction when the 20 most important nodes (based on betweenness, closeness, degree, and eigenvector centrality measures) are removed from the network (2010, 2011, and 2012) and random node removal. Diamond symbols are nodes identified as important based on each centrality metric.

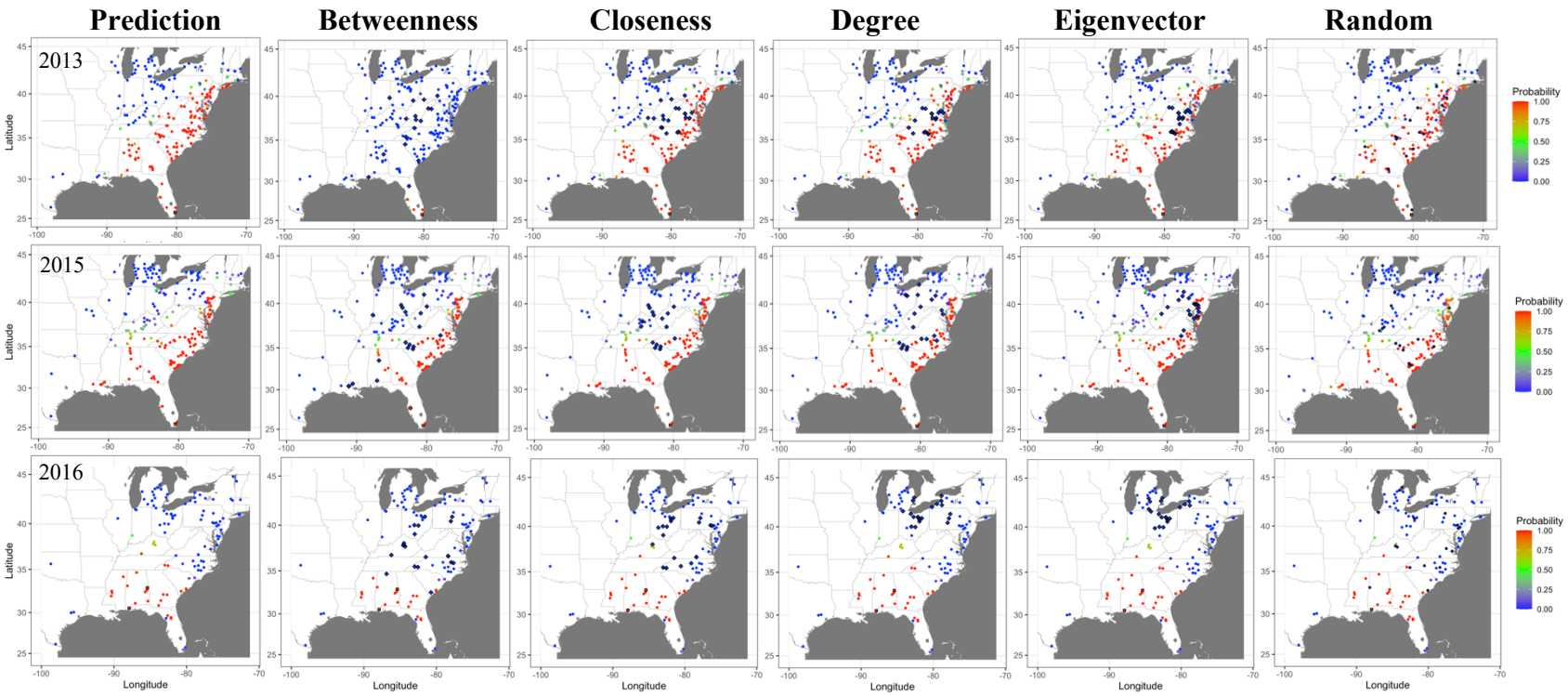


Figure S3.17. Prediction of cucurbit downy mildew outbreaks in the eastern United States by week 25 for all nodes present in the network (i.e., prediction) compared to prediction when the 20 most important nodes (based on betweenness, closeness, degree, and eigenvector centrality measures) are removed from the network. (2013, 2015, and 2016). Diamond symbols are nodes identified as important based on each centrality metric.

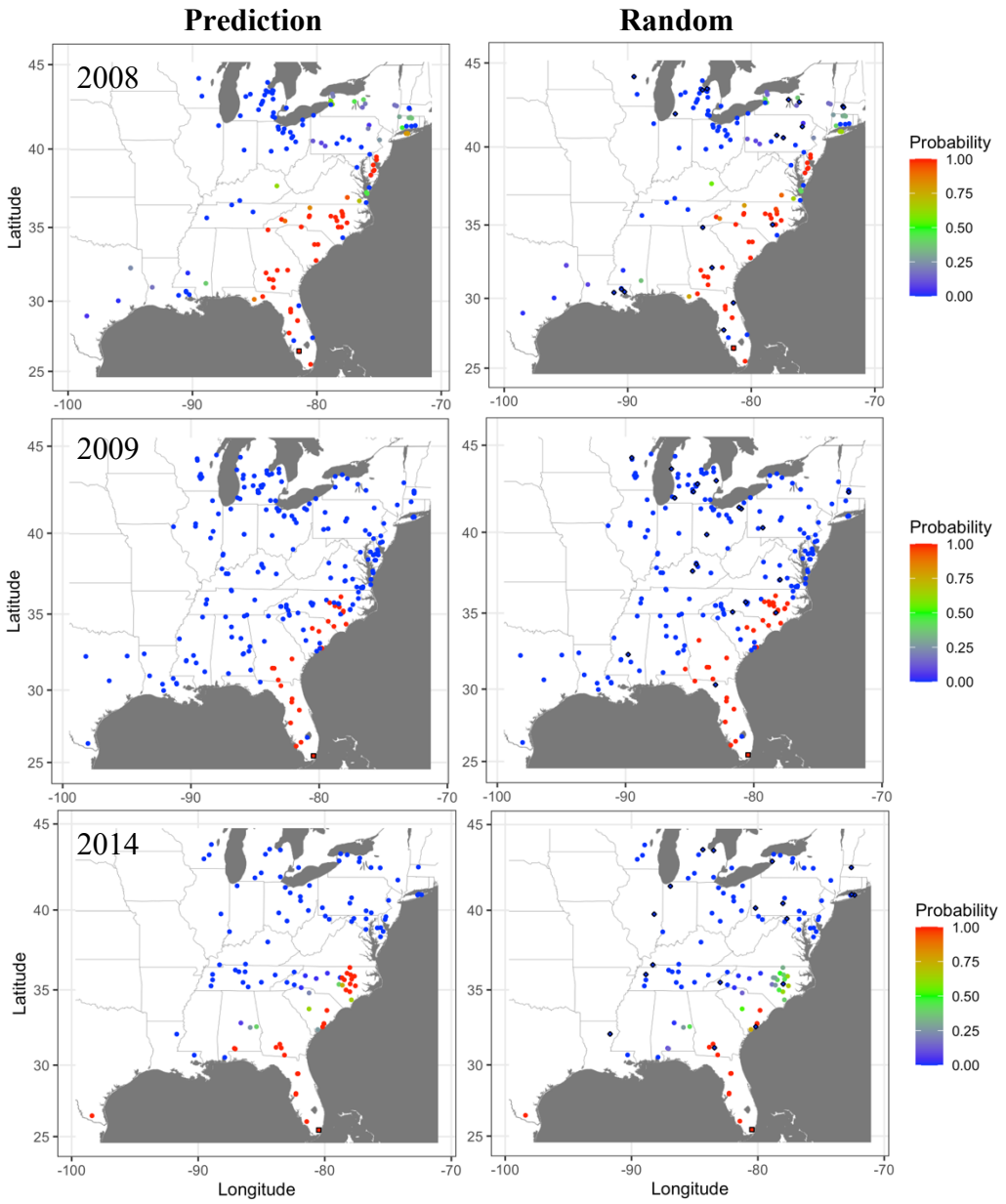


Figure S3.18. Prediction of cucurbit downy mildew outbreaks in the eastern United States by week 25 for all nodes present in the network compared to a prediction when 20 random nodes are removed from the network (2008, 2009, and 2014). Diamond symbols are nodes identified as important based on each centrality metric.

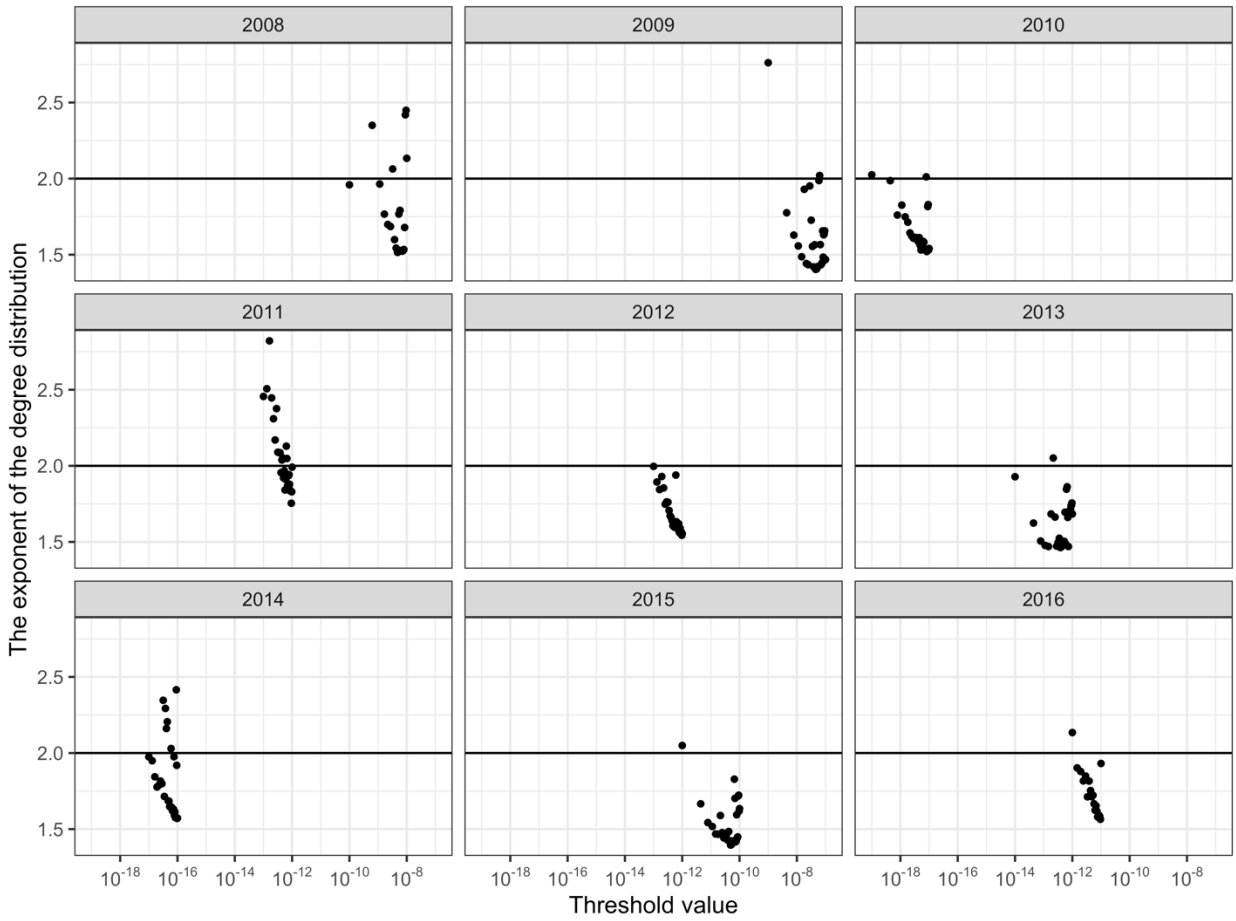


Figure S3.19. The exponent of the degree distributions for 2008 to 2016 networks created using different thresholds. A value of 2 indicates that a network is scale-free, i.e., the degrees follow a power-law distribution.

CHAPTER 4

A General Framework for Spatiotemporal Modeling of Epidemics with Multiple Epicenters: Application to an Aerially Dispersed Plant Pathogen

Submitted to *Frontiers in Applied Mathematics and Statistics*

A.M.E Ojwang[1], T. Ruiz[2], S. Bhattacharyya[2], S. Chaterjee[3], P.S. Ojiambo[4] and D.H. Gent[2]

[1] Biomathematics Program, Department of Mathematics, North Carolina State University, Raleigh, NC, USA

[2] Department of Statistics, Oregon State University, Corvallis, OR 27695, USA

[3] Department of Mathematics, City University of New York, New York City, NY, U.S.A.

[4] Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, U.S.A.

[5] U.S. Department of Agriculture, Agricultural Research Service, Corvallis, OR, U.S.A.

CONTRIBUTION

As the first co-author of this publication, I was responsible for coding and running the simulation study of the models presented, analyzing the simulation results, preparing the simulation figures, writing the abstract, introduction, results, and proofreading the manuscript.

Abstract

The spread dynamics of long-distance-dispersed pathogens are influenced by the dispersal characteristics of a pathogen, anisotropy due to multiple factors, and the presence of multiple sources of inoculum. In this research, we developed a flexible class of phenomenological spatiotemporal models that extend a modeling framework used in plant pathology applications to account for the presence of multiple sources and anisotropy of biological species that can govern disease gradients and spatial spread in time. We use the cucurbit downy mildew pathosystem (caused by *Pseudoperonospora cubensis*) to formulate a data-driven procedure based on the 2008 to 2010 historical occurrence of the disease in the U.S. available from standardized sentinel plots deployed as part of the Cucurbit Downy Mildew ipmPIPE program. This pathosystem is characterized by annual recolonization and extinction cycles, generating annual invasions at the continental scale. The data-driven procedure is amenable to fitting models of disease spread from one or multiple sources of primary inoculum and can be specified to provide estimates of the parameters by regression methods conditional on a function that can accommodate anisotropy in disease occurrence data. Applying this modeling framework to the cucurbit downy mildew data sets, we found a small but consistent reduction in temporal prediction errors by incorporating anisotropy in disease spread. Further, we did not find evidence of an annually occurring alternative source of *P. cubensis* in northern latitudes. However, we found a signal indicating an alternative inoculum source on the western edge of the Gulf of Mexico. This modeling framework is tractable for estimating the generalized location and velocity of a disease front from sparsely sampled data with minimal data acquisition costs. These attributes make this framework applicable and useful for a broad range of ecological data sets where multiple sources of disease, or other organisms, may exist and whose subsequent spread is directional.

1. Introduction

Epidemics caused by invasive pathogens can be managed through several approaches that include quarantine, containment, eradication programs, and chemical control measures. Understanding the risk of disease invasion is vital in facilitating the planning of disease control, prediction and prevention of epidemics, and development of mitigation policies (Ojiambo et al., 2017). These needs are particularly acute for fecund organisms capable of long-distance dispersal that are not spatially restricted. Dispersal of these organisms is a fundamental process with many implications for invasion ecology. The characteristics and frequency of long-distance dispersal may influence processes such as spatial distribution of an organism, gene flow between populations, pathogen population expansion, and invasiveness (Clark et al., 2001; Ibrahim et al., 1996; Kot et al., 1996; Severns et al., 2019; Wingen et al., 2007). Dispersal characteristics of a pathogen are also central to formulating sound policies for mitigation of ensuing epidemics, such as predicting the first appearance of disease and timing of intervention efforts (Severns et al., 2019).

Diverse disease organisms affecting plants, animals, and humans, may generate patterns of disease due to long-distance dispersal that can be explained by similar models provided that inoculum moves over long distances (Mundt et al., 2009a and 2009b). Plant disease epidemics, therefore, are excellent model systems for understanding dispersal and its determinants due to the annual occurrence of epidemics and experimental tractability of the systems. One such disease example is cucurbit downy mildew, caused by the oomycete *Pseudoperonospora cubensis*. Cucurbit downy mildew is a major concern for growers in the eastern USA leading to substantial economic losses. For example, in 2004 alone, the epidemic of cucumber resulted in 16 million USD economic loss (Colucci et al., 2010).

In the U.S. and central Europe, the pathogen exhibits annual recolonization and extinction cycles, generating annual invasions at the continental scale because *P. cubensis* is aerially dispersed, and sporangia can be transported long distances (Jaing et al., 2020; Ojiambo et al., 2011). Additionally, *P. cubensis* is an obligate parasite that must overwinter on living host tissue. In the U.S., this is thought to restrict overwintering under natural conditions to frost-free areas below approximately 30-degree latitude (Ojiambo and Holmes, 2011; Ojiambo et al., 2015). Historical data on the occurrence of the disease is available from standardized sentinel plots deployed as part of the Cucurbit Downy Mildew ipmPIPE program (Ojiambo et al., 2011b). Furthermore, the disease is economically important and can result in complete crop loss in the absence of adequate control measures (Cohen et al., 2015; Ojiambo et al., 2015). Successful management also requires that control measures be implemented just before or at the first detection of the disease in a field or region.

Simple predictive models with analytical solutions have been used to analyze disease spread in plant epidemics when mechanistic models do not exist. We consider phenomenological models with empirical support in plant disease epidemiology as starting points for our framework. We focus on widely used models for both the temporal and spatial behavior of pathosystems driven by aerial dispersal. Infection of cucurbits by *P. cubensis* results in epidemics where inoculum is produced by plants previously infected during the same epidemic that season. In plant disease epidemiology, such epidemics are termed polycyclic, and the logistic model is one of the simplest in a class of models that accurately approximate the behavior of these epidemics over time. The model represents the rate of change in disease intensity over time as proportional to the disease intensity y, and the healthy quantity $1 - y$

$$\frac{dy}{dt} = ay(1 - y) \quad (1)$$

where a is the rate of disease progression. The observed disease on individuals infected during the epidemic is represented by $y = Y/N$, where Y represents the disease in absolute units such as the number of lesions, infected leaves or plants, and N is the total number of individuals or plant area that can be infected. The value of y is bounded between 0 and 1, inclusively. For an epidemic to occur, there must be contact between inoculum and disease-free individuals. The latter is incorporated into the model by the expression $1 - y$. Production and dispersal of inoculum from infected individuals, infection of healthy individuals, and subsequent production of new inoculum by the newly diseased individuals are incorporated into the model by the rate parameter a (Madden et al., 2007). This model framework is widely used in plant disease epidemiology to describe diverse pathosystems.

Pathogens exhibiting long-distance dispersal result in epidemics with accelerating velocity over time that are often difficult to control (Severns et al., 2019); inoculum of such pathogens arises from an initial disease focus (or multiple foci) and travels long distances where it may cause disease far from the initial focus. The long-distance spread of disease generates a spatial dispersal gradient relative to the focus - the rate of decrease in inoculum density with distance from a source (Gregory 1968). For aerially dispersed pathogens, the wind is the main dispersal mechanism of inoculum. Epidemics driven by aerial dispersal exhibit wave-like behavior in which spatial dispersal at any given time can be accurately approximated by the power law (Ojiambo et al., 2017). The power-law model is of the form

$$\frac{dy}{dt} = \frac{-by}{r} \tag{2}$$

or a modified version

$$\frac{dy}{dt} = \frac{-by}{r + \lambda} \tag{3}$$

where r is the maximum distance of disease spread (at time t) expressed as a radius from the epicenter, b is the spread parameter (unitless) and λ is a offset parameter incorporated into the model to permit calculations at $r = 0$. The power-law model only approximates epidemic behavior well on certain spatial scales; for large y and small r, these two versions of the model can produce extreme $\frac{dy}{dr}$ values inconsistent with realistic dispersal behavior for values of b and λ that approximate dispersal well farther from the source position. In addition, the model implies an upper limit to disease intensity at any given location. A simple modification overcomes these limitations:

$$\frac{dy}{dt} = \frac{-by(1-y)}{r+\lambda} \tag{4}$$

Equation (4) is known as a power-logistic model (Madden et al., 2007). This power-logistic model is consistent with empirical observations for disease spread at multiple spatial scales (Madden et al., 2007).

Disease epidemics are dynamic population processes occurring in both time and space; the above phenomenological models can jointly approximate such spatiotemporal dynamics. For sparse observational data, it is often of interest to describe the epidemic wavefront - the farthest distance from a source position. To this end, we denote the disease intensity y at a maximal location $r(t)$ at a time t by

$$y(t, r(t)|\Theta) = f(t, r(t)|\Theta)$$

where $f(\cdot)$ is some continuous function describing the variation in intensity at the wavefront over time relative to a source position given a vector Θ of population parameters. The parameters Θ characterize the spatiotemporal dynamics. Assume that the disease intensity function has derivatives

$$\frac{dy(t)}{dt} = ay(t)\big(1 - y(t)\big) \tag{5}$$

$$\frac{dy(t)}{dr(t)} = \frac{-by(t)\big(1 - y(t)\big)}{r(t) + \lambda} \tag{6}$$

where $y(t)$ is short for $y(t, r(t)|\Theta)$ and the population parameters are $\Theta = (a, b, \lambda)$. Considered jointly, these derivatives for the (spatial) power-logistic and (temporal) logistic phenomenological models characterize a broad class of potentially quite complex intensity functions f. An instantaneous measure of the epidemic velocity $v(t)$ can be expressed in terms of the maximal distance $r(t)$ from the epicenter at time t as

$$v(t) = \frac{dr(t)}{dt} = \frac{dr(t)}{dy(t)} \times \frac{dy(t)}{dt} = \frac{-a(r(t) + \lambda)}{b} \tag{7}$$

Ojiambo et al. (2017) used this power-logistic model to estimate the spread parameter b of epidemic waves resulting from an assumed isotropic spread of cucurbit downy mildew in the eastern U.S. They assumed that all epidemics are first observed at the same initial distance of spread r_0 (Ojiambo et al., 2017) given that *P. cubensis* overwinters in south Florida and the inoculum is aerially dispersed northward when the environment is conducive (Holmes et al., 2015; Ojiambo et al., 2015).

These spatiotemporal models implicitly assume isotropic spread since the derivatives (Equations 5 and 6) do not depend on direction from the source position spread (Mundt et al., 2009a and 2009b). However, dispersal is generally anisotropic for long-distance dispersed pathogens. Anisotropy may be due to landscape features (Taylor et al., 1993), host availability (Margosian et al., 2009), and weather, of which wind is particularly relevant for aerially dispersed organisms (Gregory, 1968). Various studies have developed anisotropic dispersal kernels to describe relatively short distance dispersal of seeds, pollen, and pathogen propagules (van Putten

et al., 2012). Inoculum dispersed in different directions from a source can be expressed in terms of either density or distance. In terms of density, anisotropy is the mean number of spores deposited in a given direction, while in distance, it is the mean distance transversed by a spore in a given direction. An example is a work by Soubeyrand et al. (2007) on yellow rust of wheat caused by *Puccinia striiformis* where two anisotropy functions were explored to quantify and differentiate anisotropy in density and distance using parametric and nonparametric approaches. The nonparametric approach was used to determine the main directions and the shapes of the anisotropy functions, but without explicit linkage to covariates such as wind speed and direction. Similarly, Rieux et al. (2014) examined a range of dispersal kernels and found that disease gradients for ascospores and conidia of *Mycosphaerella fijiensis* were best described by a fat-tailed exponential power kernel and a thin-tailed dispersal kernel, respectively. Rieux et al. (2014) further estimated anisotropy in both density and distance and showed that anisotropy was correlated with averaged daily wind gust for conidia, although wind covariate information was not used explicitly to estimate anisotropy in disease gradients. These modeling frameworks incorporate anisotropy into disease gradients observed for the special case of a single pathogen generation or dispersal event but do not consider anisotropy in epidemic spread in time (Rieux et al., 2014; Soubeyrand et al., 2007; van Putten et al., 2012).

Besides anisotropy, fitting and interpreting disease gradients and dispersal are further complicated by the presence of multiple sources of inoculum (Gregory, 1968). Great care usually is taken in experimental settings to minimize background inoculum that can confound interpretation of disease gradients (Cowger et al., 2005; Rieux et al., 2014; Soubeyrand et al., 2007). Controlling for multiple inoculum sources in natural epidemics is much more complicated (Waggoner, 1962). Process-based models may be most useful in these situations for the description

and prediction of epidemics (Meyer et al., 2017), but such approaches are highly resource-intensive, and few exist in practice. A more common situation, especially with invasive organisms, is that physical process models are not yet available and resource limitations result in relatively sparse sampling and data. Thus, simpler phenomenological models are needed to derive generalized estimates of potential disease spread and probable sources of primary inoculum (Ojiambo et al., 2017).

Returning to the motivating example of cucurbit downy mildew, although *P. cubensis* may overwinter on susceptible hosts in temperature regions, an alternative source of inoculum may exist in protected cultivation (Cohen et al., 2015; Ojiambo et al., 2015; Savory et al., 2011) or potentially oosporic inoculum (Thomas et al., 2017). This hypothesis of alternative sources of inoculum has been proposed several times but never demonstrated conclusively. Thus, the cucurbit downy mildew system also may be suitable for formulating models that account for multiple sources of the initial inoculum.

In this study, we extend the work of Ojiambo et al. (2017) and Rieux et al. (2014) with a modified power-logistic model that includes anisotropy of disease in space and time and also consider multiple sources of primary inoculum. We present a flexible and generalize framework that accounts for multiple sources of inoculum and applies to cucurbit downy mildew. This flexible framework can also be extended to any pathosystem where the special conditions of isotropic spread or a single inoculum source may be too restrictive.

2. Methods

Modeling approach

Our work develops a generalization of the spatiotemporal model given by Equations 5 and 6 that modifies the power-logistic model for spatial dynamics (Equation 6) by parametrizing λ as a function that depends on the direction from the epicenter, and we apply this model framework in an analysis of cucurbit downy mildew disease data. In this section, we first present the model framework and discuss estimation. Following this, we describe the data sets analyzed. We then present application-specific details involved in our analysis. Lastly, we present a simulation study to understand the sensitivity of the modeling framework and estimation procedure to sample sizes, error variance, and aspects of epidemic behavior.

2.1. Anisotropic multi-source velocity model

For the purpose of exposition, our model is first presented with reference to a single source and then extended to describe simultaneous dispersal from multiple sources by introducing a latent factor that indicates causal attribution to one of the sources in the model. Following a description of the multi-source extension, we present an iterative estimation method based on the Expectation–Maximization (EM) algorithm.

2.1.1. Single source model

First consider a model for disease emanating from a single source point located in two-dimensional Cartesian space, with the source location denoted $x_0 \in R^2$. Let $(r(t), \phi)$ denote the polar coordinates associated with the maximum distance $r(t)$ in direction ϕ of the disease wave

front relative to x_0 at time t. Let the parametric model for disease intensity at location $(r(t), \phi)$ and time t be denoted by

$$y = y(t, r(t), \phi|\Theta) = f(t, r(t), \phi|\Theta) \tag{8}$$

Let the intensity function have derivatives

$$\frac{dy}{dt} = ay(1-y) \tag{9}$$

$$\frac{dy}{dr(t)} = \frac{-by(1-y)}{r(t) + g(\phi)} \tag{10}$$

where $g(\phi)$: $[0,2\pi] \rightarrow R$ is a function of the angle between the location x and the source point x_0, and a and b are parameters of the model. The function g induces spatial anisotropy by allowing the rate of change of disease incidence with distance from the source to depend on direction.

The explicit form for $f(.)$ is obtained by integration with respect to r given the boundary condition that for the differential equations 9 and 10 at $t = t_0 > 0$ for each angle ϕ, $y(t_0, r(t_0), \phi) = y_0(\phi)$ and $r(t_0) = r_0$. The value of $y_0(\phi)$ may vary depending on the source and the epidemic under consideration. First integrating Equation 10 for a fixed ϕ gives

$$\log\left(\frac{y}{1-y}\right) = -b \log\left(1 + \frac{r}{g(\phi)}\right) + c(\phi) \tag{11}$$

where $c(\phi)$ is a constant of integration for fixed ϕ. Then, integrating Equation 11 for a fixed ϕ gives

$$\log\left(\frac{y}{1-y}\right) = at + c(\phi) \tag{12}$$

where $c(\phi)$ is a constant of integration as in Equation 11. Now, together Equations 11 and 12 imply the generic functional form

$$\log\left(\frac{y}{1-y}\right) = -b' \log\left(1 + \frac{r}{g(\phi)}\right) + a't + c(\phi) \tag{13}$$

that satisfies jointly the differential Equations 9 and 10. Note that, b' and a' are obtained such that the right-hand side of Equation 13 is a convex combination of the right-hand side terms in Equations 11 and 12 (one example is given by $b' = \frac{b}{2}$ and $a' = \frac{a}{2}$). Algebraic rearrangement of Equation 13 yields that the explicit form for $f(.)$ up to constants of integrations for fixed ϕ is

$$y(r, \phi, t|\Theta) = \frac{1}{1 + \left(1 + \frac{r(t)}{g(\phi)}\right)^{b} \exp(-at)\, A(\phi)} \tag{14}$$

where the parameters are $\Theta = \left(a, b, g(\cdot)\right)$ and $A(\phi) = \frac{1}{\exp(c(\phi))}$. This is a spatiotemporal process for disease intensity with spatial kernel $F(\phi)\left(1 + \frac{r(t)}{g(\phi)}\right)^{-b}$ in the disease wave front. We note that this result is a generalization of the 'geometric' spatial kernel considered in Rieux et al. (2014) among the candidate models for anisotropic dispersal densities, wherein the anisotropy-inducing function g is a radial density; Rieux et al. (2014) consider the Von Mises distribution for a specific functional form.

A derived model for velocity describes the movement of an epidemic wavefront. As noted in the introduction, this can be especially useful for epidemiological data that are sparse in space and time, which contain relatively less information about the spatiotemporal distribution of disease incidence. From Equations 9 and 10, velocity is given by

$$v = -\frac{dr}{dt} = M\left(\frac{1}{g(\phi) + r}\right)^{-1} \tag{15}$$

where $M = \frac{a}{b}$. Integrating 15 yields

$$\log\left(1 + \frac{r}{g(\phi)}\right) = -Mt + h(\phi) \tag{16}$$

where $h(\phi)$ is a normalizing constant for fixed angle ϕ. We note that Equation 16 is linear in time and can be fit to obtain estimates of M, h, and g using regression methods (as described in further detail below).

2.1.2. Multiple source model

We extend the velocity model above to describe epidemics emanating from K source points. A summary of the notations used is given in Table 4.1. Let $\left(x^{(1)}, \dots, x^{(K)}\right)$ denote the source locations; for each $k = 1, \dots, K$, $x^{(k)} \in \mathbb{R}^2$. Now an arbitrary location $x \in \mathbb{R}^2$ is associated with K sets of polar coordinates $\left(r^{(1)}, \phi^{(1)}\right), \dots, \left(r^{(K)}, \phi^{(K)}\right)$, where the kth polar coordinate pair indicates the distance $r^{(k)}$ and angle $\phi^{(k)}$ to the kth source point $x^{(k)}$. A depiction of this data representation is given in Fig 4.1. Applying the model framework above to each set of coordinates yields the collection of velocity models

$$\log\left(1 + \frac{r^{(k)}}{g_k(\phi^{(k)})}\right) = -M_k t + h_k\left(\phi^{(k)}\right), \qquad k = 1, \dots, K \tag{17}$$

Now, if multiple sources are present, any given location could be subject to disease exposure from as many as K wavefronts moving simultaneously. Yet, depending on conditions, the movement patterns of the wavefronts, and relative distances to each epicenter, an infection event at any particular time and location is attributable to the different sources with varying probability. In other words, disease at particular locations is more likely due to certain sources rather than others. To accommodate this intuition, a latent process Z is introduced that indicates the relative probabilities of disease associated with each of the K sources, and the collection of models given in Equation 17 describe (r, ϕ, t) conditional on the possible values of Z.

$$Z \sim \text{Multinomial}\left(1, \left(p^{(1)}, \dots, p^{(K)}\right)\right) \tag{18}$$

For example, $P(Z = 1) = p^{(1)}$ indicates that an infection event is caused by source 1 with probability $p^{(1)}$. We then assume that a disease occurrence is described by each of the K velocity models given in Equation 17 with probabilities $p^{(1)}, \ldots, p^{(K)}$. That is, for an arbitrary disease occurrence at time t, we posit the set of wave front descriptions

$$\log\left(1 + \frac{r^{(k)}}{g_k(\phi^{(k)})}\right) = -M_k t + h_k\left(\phi^{(k)}\right) \tag{19}$$

with probability $p^{(k)} = P(Z = k)$ for $k = 1, \ldots, K$. This framework makes the implicit assumption that disease is caused by inoculum produced at exactly one source. However, it will be seen that our estimation method does not involve a hard classification rule for disease observations and thus we instead specify observation weights for each velocity model according to estimated probabilities $p^{(1)}, \ldots, p^{(K)}$.

2.1.3. Estimation

We propose an estimation procedure wherein velocity models are fit using regression methods conditional on known g_k. The functions g_k introduce anisotropy in the model by imposing directional variation in the spatial rate of change of disease incidence via the partial differential equation in Equation 10. In many applications, known variables drive anisotropy, so it is often plausible to estimate g_k from covariate information or secondary data sources.

The velocity models (Equation 17) are fitted conditional on g_k to disease occurrence data (presence or absence) of the form $\{(r_i^{(1)}, \phi_i^{(1)}), \ldots, (r_i^{(K)}, \phi_i^{(K)}), t_i\}_{i=1}^n$ indicating the locations and times of the first observed disease case. For the purpose of exposition, suppose one is fitting only

the kth model: consider just the data $\left(r_i^{(k)}, \phi_i^{(k)}, t_i\right)$ and assume $P(Z_i = k) = 1$. Now, adding an offset c_k and Gaussian error term $\epsilon_i^{(k)}$ to Equation 17 yields the statistical model

$$\log\left(1 + \frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) = c_k - M_k t_i + h_k\left(\phi_i^{(k)}\right) + \epsilon_i^{(k)} \quad \begin{cases} \epsilon_i^{(k)} \overset{iid}{\sim} N(0, \sigma_k^2) \\ i = 1, \ldots, n \end{cases} \tag{20}$$

Estimates of c_k, M_k and h_k are easily computed using semiparametric regression. Let $s_1(\cdot), \ldots, s_B(\cdot)$ denote a set of B basis functions. Now, rewriting Equation 20 we obtain

$$\log\left(1 + \frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) = c_k + (-M_k)t_i + \beta_1^{(k)} s_1\left(\phi_i^{(k)}\right) + \cdots + \beta_B^{(k)} s_B\left(\phi_i^{(k)}\right) + \epsilon_i^{(k)} \tag{21}$$

Ordinary least squares (OLS) solution to Equation 21 subsequently yields estimates of $\hat{c}_k, \widehat{M}_k$ and $\hat{h}_k = \sum_b \hat{\beta}_b^{(k)} s_b$.

Finally, this estimation strategy is extended to the full collection of K models by accounting for the latent variables Z_i that attribute each of the i data points to one of the K sources. Formally, the joint likelihood of the data arising from Equations 18 and 19 is maximized with respect to the parameters $p^{(k)} \in \mathbb{R}^N$, $\beta_k \in \mathbb{R}^{B+2}$, and σ_k^2 for $k = 1, \ldots, K$. The Expectation-Maximization (EM) algorithm is used to iteratively update estimated multinomial probabilities $\hat{p}_i^{(1)}, \ldots, \hat{p}_i^{(K)}$ for each data point in alternation with fitting the regression models in Equation 21 using the estimate $\hat{p}_i^{(k)}$ as a regression weight for the ith data point in fitting the kth model. In detail, the iterations are given by:

1. Initiate $\hat{p}_i^{(k)}$ as the weight of ith data-point to be associated with kth source, where $\sum_{k=1}^{K} \hat{p}_i^{(k)} = 1$.

2. Compute/update the estimates $\left(\hat{c}_k, \widehat{M}_k, \hat{h}_k, \hat{\sigma}_k^2\right)_{k=1}^{K}$ by fitting each of the models in Equation 21 with weights $\hat{p}_i^{(k)}$ for the ith data point and the kth model.

3. Update $\hat{p}_i^{(k)}$ by

$$\hat{p}_i^{(k)} = \frac{\varphi\left(\hat{c}_k - \widehat{M}_k t_i + \hat{h}_k\left(\phi_i^{(k)}\right), \hat{\sigma}_k^2\right)\hat{p}_i^{(k)}}{\sum_{k=1}^{K}\varphi\left(\hat{c}_k - \widehat{M}_k t_i + \hat{h}_k\left(\phi_i^{(k)}\right), \hat{\sigma}_k^2\right)\hat{p}_i^{(k)}} \tag{22}$$

where $\varphi(x, \sigma^2)$ is the probability density function of a Gaussian random variable with mean zero and variance σ^2 evaluated at value x.

4. Repeat steps 2-3 until convergence.

A simple heuristic for the initialization step is to use as $\hat{p}_i^{(k)}$ the estimated probabilities obtained by logistic regression of an indicator of whether the kth source is closest on the variables $r^{(1)}/\hat{g}_1\left(\phi^{(1)}\right), \ldots, r^{(K)}/g_k\left(\phi^{(K)}\right)$. We note that an isotropic model with one or many sources can be recovered within this framework as a special case by fixing $g_k(x) = 1/2\pi$ for $x \in [0,2\pi]$, with the consequence that $h_k \equiv 0$. The details on the derivation and explanation of the fitting procedure are given in the Supporting Information EM Algorithm.

2.1.4. Spatial and temporal predictions

Estimated models - the K models in Equation 21 - directly yield fitted values for the quantity $\log\left(1 + \frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right)$. Since this quantity does not have meaningful units, estimated times of disease occurrence conditional on location and estimated distances of occurrences from sources conditional on time and direction for each data point provide more interpretable assessments of fit quality with biological relevance. These spatial and temporal estimates are

$$\hat{t}_i^{(k)} = \frac{1}{\widehat{M}_k}\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) - \hat{h}_k\left(\phi_i^{(k)}\right) - \hat{\beta}_0^{(k)}\right) \quad (23)$$

$$\hat{r}_i^{(k)} = g_k\left(\phi_i^{(k)}\right)\left(\exp\left\{\hat{\beta}_0^{(k)} + \widehat{M}_k t_i + \hat{h}_k\left(\phi_i^{(k)}\right)\right\} - 1\right) \quad (24)$$

Since the model includes estimated probabilities that the ith data point is associated with each source (the estimates $\hat{p}_i^{(1)}, \ldots, \hat{p}_i^{(K)}$), a simple heuristic (note that this approach doesn't include a covariance term) for selecting a single temporal estimate from $\hat{t}_i^{(1)}, \ldots, \hat{t}_i^{(K)}$ and a single spatial estimate from $\hat{r}_i^{(1)}, \ldots, \hat{r}_i^{(K)}$ is to choose the estimates $\hat{t}_i^{(k)}$ and $\hat{r}_i^{(k)}$ associated with the most probable source. That is, let

$$(\widehat{t_i}, \widehat{r_i}) = \left(\widehat{t_i^{(k^*)}}, \widehat{r_i^{(k^*)}}\right) \text{ where } k^* = \text{argmax}_k\{\hat{p}_i^{(k)}\} \quad (25)$$

Then, a fitted model can be evaluated according to the spatial and temporal root mean square error (RMSE) metrics

$$\text{rmse}_t \overset{\text{def}}{=} \left(\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2\right)^{1/2} \quad (26)$$

$$\text{rmse}_r \overset{\text{def}}{=} \left(\frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2\right)^{1/2} \quad (27)$$

2.2. Cucurbit downy mildew data

Epidemics of cucurbit downy mildew recorded in the U.S. from 2008 to 2016 were obtained from the data submitted to the Cucurbit Downy Mildew ipmPIPE program (http://cdm.ipmpipe.org). The ipmPIPE is an information and decision support system that gathers pertinent data (disease occurrence in cucurbit production areas), applies predictive models to the data, incorporates expert interpretation, and communicates near-real-time output to cucurbit growers, extension personnel, crop consultants (Ojiambo et al., 2011). Records of outbreaks in the

system include disease reports from a network of regularly monitored sentinel plots as well as voluntary disease reports from non-sentinel plots submitted by commercial growers, agricultural researchers, and the general public. We describe below the two types of disease reports and a subset of the data selected for analysis.

Sentinel plots were fixed locations planted with different cucurbit host types for monitoring downy mildew outbreaks and strategically placed within specific states at locations that collaborators can easily access. During the years 2008-2016, the sentinel plots were located at research facilities or in commercial fields with standard dimensions of 15 m x 61 m and were georeferenced using the Global Positioning System. These plots were monitored for disease symptoms weekly to biweekly by cooperating scientists and extension specialists and were planted with susceptible, early maturing cultivars. The cucurbit host types grown in the sentinel plots were *Cucumis sativus* (cucumber cv. Straight 8 and Poinsett 76), *Cucumis melo* (cantaloupe cv. Hales Best Jumbo), *Cucurbita pepo* (acorn squash cv. Table Ace), *Cucurbita maxima* (giant pumpkin cv. Big Max), *Cucurbita moschata* (butternut squash cv. Waltham), and *Citrullus lanatus* (watermelon cv. Micky Lee) (Ojiambo et al., 2011). The compiled data set on sentinel plot disease reports consist of the date of first observed occurrence of disease, the reporting date, affected host type, the incidence of plants affected, and plot location.

Cucurbit downy mildew was also monitored via voluntary reporting in locations not designated for regular surveillance. These locations include commercial fields, research plots, and home gardens. Compiled data on voluntary reports consisted of the date of first observed occurrence of disease, the reporting date, location, and affected host type (if provided). This information is potentially instructive for understanding the distribution and appearance of cucurbit downy mildew, but subject to greater uncertainty with respect to the timeliness of disease detection

due to the non-standardized nature of the plant populations, potential confounding from fungicide applications, and the absence of regular monitoring and reporting protocols.

We selected a subset of the disease reports from which to model epidemic wavefronts using the framework described above. The sub setting strategy was intended to capture a single wave as best as possible while ensuring uniform reliability on the timeliness of reports. First, for the reliability of timeliness, we considered only sentinel reports. This was thought to better ensure consistent variation across reports in the accuracy of dates of first observed disease occurrences due to a fixed observation frequency and protocol. Second, late-season outbreaks are known to occur due to later-planted cucurbit crops that are common in southern and mid-Atlantic regions of the U.S. (Ojiambo et al., 2017). Thus, we sought to capture the first outbreak each year by restricting attention to reports in which the date of observed occurrence is before August. Finally, we selected data from 2008, 2009, and 2010 to capture annual variation, and chose these specific years due to a relatively greater number of sentinel plots available. From the resulting reports, we compiled data on the location, date of symptom onset (presence of disease at any level), and host type from each report.

2.3. Application details

The three consecutive years of selected sentinel reports were analyzed separately by fitting isotropic (I) and anisotropic (A) one-source (OS) and two-source (TS) velocity models to data from each year. In order to apply the model framework to this specific dataset, we identified potential source locations from an exploratory analysis of early occurrences and developed a simple method of estimating the functions g_k from meteorological information known to drive dispersal.

2.3.1. Selection of source locations for cucurbit downy mildew data

Source locations were specified as county centroids. To identify putative source locations for each year, we examined both sentinel and voluntary reports of early disease occurrences for geographical location and timing. The first observation of disease occurrence reliably in southern Florida in every year, so the centroid of the county in which the first disease symptoms were reported each year was fixed as the main source point. In addition, early occurrences are often observed in the southwestern United States and the Great Lakes region before expected dispersal from the source point in Florida. We identified several counties in northern latitudes (Erie and Wayne counties in Ohio, and Niagara County in New York) that had early occurrences in multiple years, and several counties in the southwestern region (Brazos and Hidalgo counties in Texas, Vernon County in Louisiana, and Payne County in Oklahoma) that had early occurrences in multiple years. We considered each of these counties as possible locations for a second source each year. Based on the reports in each region, the earliest disease occurrences were used to identify dates at which a putative source in each region might appear. We note here that the alternate sources specified are not necessarily the actual location of overwintering of *P. cubensis* but are a reasonable proxy for an alternate source of inoculum when placed within the path of the wave-front emanating from the true source.

2.3.2. Estimation of g_k from meteorological data

We estimated the functions g_k based on meteorological data measured at the source locations since variation in wind direction and speed are the primary drivers of anisotropic dispersal. Hourly wind direction and speed near each county centroid with a sentinel plot or imputed disease source was derived from weather observations in the National Oceanic and

Atmospheric Administration Integrated Surface Database (Smith et al., 2011) and were provided by BASF (Research Triangle Park, Raleigh, NC).

Nonparametric kernel density estimates of the functions g_k were computed from the collection of hourly wind directions at each of the k sources over the time interval represented in the disease data. If $\theta_i^{(k)}$ denotes the angle of the predominant wind direction at time point i and source location k, the wind direction data $\theta_1^{(k)}, \ldots, \theta_n^{(k)}$ is treated as a sample of size n_k on the unit circle centered at the source point $x^{(k)}$. For each k, we computed a kernel density estimator $\hat{g}_k$ of the form

$$\hat{g}_k(x) = \frac{C(h)}{n} \sum_{i=1}^{n_k} K\left(\frac{1 - x\theta_i^{(k)}}{h^2}\right), \quad x \in [0,2\pi] \tag{28}$$

where h is a positive number and $C(h)^{-1} = \int_0^{2\pi} K\left((1 - x\theta)/h^2\right)d\theta$ is a normalizing constant. For the application in this work, the kernel function $K(z) = \exp - z$ is used.

2.3.3. Application of model framework

To apply our modeling framework in the analysis of the cucurbit downy mildew data, we calculated two alternate responses: a response for the isotropic models, $\log\left(1 + r_i^{(k)}\right)$, and a response for the anisotropic models, $log\left(1 + r_i^{(k)}/\hat{g}_k\left(\phi_i^{(k)}\right)\right)$ for each data point $i = 1, \ldots, n$, and estimated the velocity models as described above.

2.4. Simulation study

We conducted a set of simulations to quantify parameter recovery using the model and estimation procedure for different sample sizes and scenarios when disease spread was weakly to

strongly anisotropic. First, we set two source locations well separated in two-dimensional space at Cartesian coordinates of (0,0) and (2000, 2000) representing the first and second sources, respectively. These locations approximate the spatial scale in kilometers of the cucurbit downy mildew sources in Florida and an alternate source of interest in the Upper Midwest. For sample sizes of n = 50, 100, 250, and 500, we fixed the proportion of disease attributable to the first and second sources as 0.7 and 0.3, respectively. We chose the Von Mises density function to generate circular normal data that could induce anisotropy in disease spread relative to the two locations. The Von Mises function has two parameters: μ, a location measure, and κ, a concentration measure, and is given by

$$f(x) = \frac{e^{\kappa cos(x-\mu)}}{2\pi I_o(\kappa)} \tag{29}$$

for any angle $x \in [\mu - \pi, \mu + \pi]$ where I_o is a Bessel function of order 0. The μ values were chosen such that disease spread from the two sources would be in opposite directions and overlapping in space by setting $\mu_1 = \frac{\pi}{4}$ and $\mu_2 = \frac{5\pi}{4}$. Values of $\kappa = 5$ and $\kappa = 2$ were chosen to produce strongly and weakly anisotropic spread, respectively. We generated two separate angle grids from the uniform distribution and used the Von Mises density function with μ and κ as noted to estimate $g(\phi)$.

We also simulated temporally synchronous and asynchronous epidemics by varying the daily time grids from the uniform distribution. The time grid ranged from the day of year 50 to 150 for synchronous spread from the two sources, and days 50 to 150 for source 1, and days 100 to 150 for source 2 for asynchronous epidemics. We then generated data according to the two models by inputting the corresponding time grid, angle grid, and coefficients according to a simplified version of the model in equation 21,

$$y_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} t_i + \beta_2^{(k)} sin(\phi_i) + \beta_3^{(k)} sin\left(\phi_i + \frac{\pi}{4}\right) + \epsilon_i^{(k)} \quad (30)$$

where, $\epsilon_i^{(k)} \sim N(0, \sigma^2)$. We consider that the error variance σ^2 is same for both the sources. We evaluated the sensitivity of the parameter estimates to error variance by varying σ^2 two-fold. The distance from each source location (r) was then calculated by back transforming (r). We then calculated the corresponding x and y Cartesian coordinates.

We pooled the simulated Cartesian coordinates and times, converted the Cartesian coordinates to sets of polar coordinates for each source, and applied the fitting procedure to estimate coefficients. We ran 1000 simulations and calculated the mean and standard deviation of the estimated coefficients, as well as the proportion of the n samples attributed correctly to each of the two sources in each of the 1000 simulated epidemics. Examples of individual simulations are presented in Figure 4.2.

3. Results

In this section, we present analyses of disease data from 2008 through 2010. For each year, several candidate models were fit. We considered both one-source and two-source models in each year, and anisotropic and isotropic versions of each. In addition, for the two-source models, we considered two alternate source locations based on the considerations discussed above.

3.1. Simulation study

The mean and standard deviation of the estimated coefficients for n = 50 and 250 for a two-source epidemic with an anisotropic spread of disease (Von Mises distribution κ = 5) are given in Table 4.2. Results for n = 100 and 500 are given in Supporting Table S4.1 and results for κ = 2 are given in Supporting Figures S4.1, S4.2, and S4.3. Parameter estimates were generally accurate

across the various sample sizes and whether epidemics at the two sources were initiated synchronously or asynchronously, but sensitive to error variance. There was a slight increase in the mean values of the intercept estimates as n increased from 50 to 250, particularly as the error variance increased. Expectedly, the standard deviation of the parameter estimates increased with error variance and decreased with n. Overall, the estimates were highly accurate in all scenarios except a large sample size with high variance and an asynchronous start (n = 250, $\sigma^2 = 1$). We hypothesize that the disease observations become more densely mixed under this setting and thus make the sources more difficult to discern, thereby compromising estimation of model parameters. We note in particular that estimates for the less dominant source (source 2) are most severely compromised, consistent with our observation below that more challenging simulation conditions tend to cause misattribution to the dominant source.

The estimated probability of disease being due to spread from one of the two sources was used to attribute disease to each of the sources in the simulations and calculate the overall mean proportion of disease correctly assigned to the true source (Figures 4.3 and 4.4). Disease due to the dominant source, source 1, was attributed correctly to this source in most simulations independent of sample size, error variance, location of the source, or other epidemic conditions specified. This behavior is consistent with expectations, as without additional information the EM algorithm attributes disease due to the most abundant, dominant inoculum source. The disease was less often attributed correctly to the less abundant second source, although source attribution here was still relatively accurate. Classification accuracy was sensitive to the placement of sources in space, diminishing as the two-source were more closely situated. Classification accuracy decreased when epidemics initiated from the two sources were temporally asynchronous, anisotropy was stronger, and error variance was larger. Although we see that some of the observations from the

less abundant source can be incorrectly allocated to the more abundant source, the estimates of the model parameters for the two sources are still relatively accurate and capture the behavior of the spread from each of the sources in terms of velocity of spread and anisotropy.

3.2. Estimation for Cucurbit Downy Mildew Dataset

Parameter estimates are reported for six models in each year: three isotropic models having one source only, an alternate source in the north, and an alternate source in the southwest; and three anisotropic models with the same source locations. The isotropic models are referred to as 'Isotropic One-Source' (IOS), 'Isotropic Two-Source (Southwest)' (ITS-SW), and 'Isotropic Two-Source (North)' (ITS-N); and similarly, the anisotropic models are referred to as 'Anisotropic One-Source' (AOS), 'Anisotropic Two-Source (Southwest)' (ATS-SW), and 'Anisotropic Two-Source (North)' (ATS-N). For the two-source models, separate velocity models are fitted corresponding to each of the two sources, and these are distinguished by indicating the source location parenthetically, *e.g.*, ITS-SW (FL) and ITS-SW (TX) indicate the two velocity models that comprise the ITS-SW model. Since one of these sources is always located in Florida in the analyses, we adopt the convention of referring to the two sources as the Florida source and the 'alternate' source.

Graphical and tabular representations of spatial and temporal prediction errors are reported for each of the models fitted to data from each year. The graphical representations focus on temporal predictions and show contours of the estimated epidemic fronts at various times. Numerical reports include spatial prediction errors, and, for the two-source models, errors from models fitted using additional alternate source locations that were considered.

These results simultaneously address several questions. First, the model comparisons

suggest whether in a given year dispersal exhibited directional variation. Second, the same comparisons provide indirect evidence for the existence of a second source, depending on whether positing such a source better explains the pattern of dispersal. Third, the prediction errors illustrate the sensitivity of results to the placement of source locations.

3.3. Disease outbreak and spread

In all the years, the disease was observed first in Florida from January to February, with reports in sentinel plots beginning in February to March (Figure 4.5). The first detection of the disease generally progressed northward with time, but with some exceptions particularly in the southwestern locations along the Gulf Coast and infrequently in the Great Lakes region such as in 2009. That is, there appeared to be anisotropy in disease spread.

3.4. Model fitting for Cucurbit Downy Mildew Dataset

The parameter estimates for one- and two-source isotropic and anisotropic models fitted to data from each of the years 2008 through 2010 are given in Tables 4.3, 4.4, and 4.5. For all years, the time parameter estimates ($-\widehat{M}_k$) for the isotropic and anisotropic one-source models (IOS and AOS) were significant (P-value), indicating that the Florida source is an important epicenter in explaining disease progression. Further, the estimates of this parameter were not affected considerably by model specification — that is, they are relatively stable across models in each year — suggesting that the epidemic velocity associated with the Florida source is a relatively strong signal in the data.

For the two-source models, estimates of the time parameter associated with the alternate source were not significantly (P-value) different from 0 when the alternate source was placed in

the Great Lakes region (ITS-N and ATS-N in all years). This indicates that the data provide little evidence of dispersal emanating from the northern alternate source location, suggesting that no epicenter was present in the region. In these cases, when the northern source is included in the model, the contribution of the associated velocity model in explaining disease progression is to generate predictions of occurrences at time-invariant distances from the source based on certain observations in the dataset. In many cases, the estimated parameter was negative, indicating a slightly contracting front toward the source. By contrast, time parameters associated with the alternate source were positive and significant (P-value) when the alternate source was placed in the southwest (ITS-SW and ATS-SW in 2008 and 2009; the estimated ITS-SW or ATS-SW models in 2010 reduced to single-source models, as all data points had low estimated probabilities of being caused by a southwestern source).

The basis parameter estimates $\hat{\beta}_1^{(k)}$ and $\hat{\beta}_2^{(k)}$, which when combined give estimates of the normalizing functions $\hat{h}_k$, are of varying significance depending on year and velocity model. In 2008, these parameters are only significant (P-value) for the ATS-SW model; in 2009, they are significant for every model except ATS-N; and in 2010, they are not significant for any model. Since the normalizing functions h_k are functions of angle, the significance of these parameters indirectly indicates the statistical strength of evidence that anisotropy is present. Thus, in 2008 there is evidence for anisotropic spread from the southwestern source, and in 2009, there is evidence of anisotropic spread from all source locations.

Finally, the estimates for the two-source models suggest that the importance of including a second source varies depending on the year and source location. None of the velocity models associated with northern source locations had significant non-intercept parameter estimates. In

2008 and 2009, the velocity model associated with the Texas source included significant terms. Yet, in 2010, the ITS-SW and ATS-SW models attributed all data points to the Florida source.

3.5. Spatial and temporal predictions

Estimated epidemic fronts from each of the six models discussed above for each of the years 2008 to 2010 are shown in Figures 4.6, 4.7, and 4.8, respectively, along with time-of-occurrence errors for each data point. Anisotropy in disease spread was apparent in the models accounting for the unequal velocity of spread in space, with the direction and magnitude varying depending on the specific source or combinations of sources (regardless of the significance of the estimates of h_k). Predicted expansion of the epidemic wavefront indicated an acceleration of epidemic velocity over time from the initial disease focus when the focus was placed in the southwest extent of the spatial domain or Florida. This was true for all years and models, consistent with the positive sign of the coefficient for the time variable associated with these models and sources ($-\widehat{M}_k$ in Tables 4.3, 4.4, and 4.5). In contrast, disease sources placed in a northern location near the Great Lakes only displayed this behavior (an expanding predicted wavefront) in 2010. In 2008 and 2009, the predicted wavefronts were either little changed over time (2008) or indicated a gravitational pull behavior (2009) due to near zero or negative parameter estimates for the time parameter (Tables 4.3 and 4.4).

Accounting for anisotropy in one source models reduced RMSE measured in time slightly (0.76 to 1.35 days) but consistently (Table 4.6; Supporting Figures 4.5 to 4.7); spatial errors were not consistently reduced in these data sets. For multiple source models, prediction errors in time and space varied over orders of magnitude depending on the model and specific year (Figures 4.6 to 4.8 and Table 4.7). Among the multiple source models, some anisotropic models reduced

temporal and spatial errors for some years as compared to the corresponding isotropic model. However, no single more complex model consistently reduced prediction errors across all years when multiple sources were included.

RMSE for multiple source models was sensitive to the placement of the alternate source in both space and time (Table 4.7). Model sensitivity to source placement was particularly acute for alternate sources placed in northern latitudes. Imputing sources in certain locations and times led to massive prediction errors in some instances, for example, when the source was placed in Niagara County, New York in 2009. Generally, reductions in RMSE in space or time were most often observed with two source models when disease spread was isotropic, and the second source was placed in the southwestern extent of the spatial domain. Conversely, two source models with the largest RMSE were most often associated with an alternate source sited in a northern latitude Table 4.7, Figures 4.6 to 4.8).

3.6. Source attribution

The modeling framework includes the estimation of the most probable source *k* resulting in disease at a distant location when multiple sources are specified. Disease outbreak was attributable to different primary sources depending on the year, location of the alternate source, and anisotropy (indicated by plotting character (Figure 4.6 to 4.8). Disease outbreaks in Florida and other southeastern states were invariably attributed to the source in southern Florida, independent of anisotropy or the specification of another source. In other regions, the source deemed most probable for disease outbreak at a given location depended on where sources were placed. Proximity was associated with whether a source was the most probable cause of disease at a given sentinel plot, but with some notable exceptions. For instance, setting a source in the Great

Lakes region led to most disease outbreaks in the Upper Midwest, Northeast, and northern mid-Atlantic region to be attributable to the northern source rather than a source in Florida. With an alternate source sited in the southwest, plots on the western and northern edge of the Gulf Coast were mostly attributed to this source, with ensuing disease spread to the northeast (Figure 4.6) or north (Figure 4.7). In 2010 there were no contours associated with the southwest source (Figure 4.8) as only two sentinel plots in Texas and Michigan were attributed to that source.

We again emphasize here that prediction errors associated with any of these models varied depending on the year and specific model and were not necessarily improved uniformly as compared to the corresponding isotropic one source model (Tables 4.6 and 4.7).

4. Discussion

We have developed a generalized, wide, and flexible class of spatiotemporal models capable of accounting for the presence of any number of initial inoculum sources and any kind of anisotropic spread of biological species that can govern disease gradients and spatial spread in time. We have also built a data-driven procedure, which selects an appropriate model from the above-mentioned class of models and provides computationally efficient estimates of the model parameters. This framework is well suited to infer the probable sources of disease spread responsible for later outbreaks at distant locations. We successfully applied this approach to predict the spread of cucurbit downy mildew in the eastern U.S., although the class of models and estimation methods are directly applicable to any disease or organism where long-distance dispersal may occur. The novelty of the class of models and estimation framework is multi-fold, as we describe below.

Previous models that describe or predict the extent of disease spread and velocity of epidemics assume dispersal is isotropic (Mundt et al., 2009a and 2009b; Ojiambo et al., 2017). This assumption usually is unrealistic because wind tends to be directional, weather gradients exist, host connectivity is patchy, inoculum source strength varies between field and regions, and landscape and terrain features influence transport and deposition of inoculum (Meentemeyer et al., 2012; Xing et al., 2020). Anisotropy may occur at multiple spatial scales, ranging from individual plants (Farber et al., 2017), individual fields (Cowger et al., 2005; Mundt and Sackett 2012), the mesoscale (Gent et al., 2019b), and the landscape or continental scale (Mundt et al., 2009; Sutrave et al., 2012). Soubeyrand et al. (2007) and Rieux et al. (2014) incorporated anisotropy into their models for describing disease gradients resulting from dispersal due to essentially one generation of plant pathogenic fungi but did not consider anisotropy in epidemics over time. The models we derived in this study accommodate both spatial and temporal components. The anisotropic model framework assumes that the rate of change of disease incidence with distance from a source depends on the direction. The rate of change with time is independent of location in the present framework but could be modified to allow the rate of change of disease incidence with distance to depend both on direction and time provided a richer data set for parameter estimation.

The importance of accounting for anisotropy will vary depending on the specific system under investigation. In the motivating example of cucurbit downy mildew used here, there was a small but consistent reduction in temporal prediction errors by incorporating anisotropy in disease spread. A reduction of multiple days is biologically relevant for aerially dispersed organisms with high reproductive potential and short generation times, where even a brief lag in implementing control measures may substantially diminish the efficacy of control measures and containment (Gent et al., 2013; Holmes et al., 2015; Severns et al., 2019). In settings where improvements in

prediction errors are inconsequential or variates related to anisotropy are unknown, an isotropic, one-source model can be recovered easily in our modeling framework as a special case.

A second novel aspect of the modeling framework derived in this study is the ability to account for multiple inoculum sources that may each produce epidemic wavefronts. Interpretation of disease gradients under natural conditions has long been recognized as a difficult process due to the potential for asynchronous and overlapping wavefronts from multiple inoculum sources (Gregory, 1968; Waggoner, 1962). The latent process introduced in our modeling framework assumes multiple sources may exist, which might better reflect conditions in natural environments when an organism is naturalized and primary inoculum is dispersed (e.g., Bergamin Filho et al., 2016; Gent et al., 2019a and 2019b; Mundt et al., 2013). The modeling framework is amenable to inference about the likelihood of disease outbreak at a specific location due to disease at multiple potential sources. This is often a basic question in invasion biology of immense importance for formulating effective management policies (Gent et al., 2019b; Graham et al., 2020), but a difficult question to address due to the stochastic nature of long-distance dispersal and technical challenges associated with its detection (Nathan et al., 2003). With multiple sources specified, our modeling approach attributes a probability to the first occurrence of the disease being associated with the specified sources. The simulation experiments indicate that the accuracy of source prediction can be influenced by the spatial proximity of the disease sources, temporal asynchrony of epidemics, the strength of anisotropy, and error variance. Source attribution is most accurate when sources are well separated in space, epidemics are temporally synchronous, and disease spread is isotropic. Source attribution error rates will increase when epidemic conditions vary in one or more of these characteristics, usually resulting in incorrect attribution of disease to the most dominant source in the landscape.

We considered examples with two sources in this work, but the approach is readily extendable to many sources provided a sufficiently dense data set for estimating the full set of models and associated latent variable process. As an example, we fit a three-source anisotropic model with epicenters placed in southern Florida, southern Texas, and a northern source in Ohio (Supporting Figure 4). The model was fit successfully, and disease in nine sentinel plots was attributed to the northern source. However, all but one of these plots were located far south of the source location. Further, the coefficient of the estimated time parameter in this model for the northern source was negative, resulting in a contracting epidemic front. In this specific example, the statistical fit of the model was improved with three sources, but the model predictions were not consistent with disease biology and ecology. However, the methodological aspects remain valid and should be suitable for other applications where sufficient data exists to avoid model overfitting.

A salient point here is that the modeling framework estimates the likelihood that the first occurrence of disease originated from a specific source but does not partition total disease intensity to one or more sources or consider later pathogen incursions. The total amount of disease at a given location can be due to multiple sources with inoculum arriving at different times. Furthermore, most disease at a location may be due to secondary or community spread following an initial infection event depending on the time since that infection and local conditions (e.g., Bergamin Filho et al., 2016; Gent et al., 2019b; Irwin, 1999). Nonetheless, understanding which source is most likely responsible for the first appearance of the disease remains highly important for understanding potential genetic founder effects, genotypic and phenotypic traits of a newly arrived pathogen causing disease, and planning mitigation strategies.

In the motivating example of cucurbit downy mildew, there is speculation and circumstantial evidence that inoculum sources outside of Florida may be important for disease in

more northern latitudes in the U.S. (Cohen et al., 2015; Ojiambo et al., 2015). Greenhouse cucumber production has been postulated as a possible alternate source of inoculum responsible for outbreaks of cucurbit downy mildew in the Great Lakes region (Holmes et al., 2015; Naegele et al., 2016; Ojiambo et al., 2015). Downy mildew can occur at damaging levels in greenhouse-grown cucurbits (Cohen et al., 2015), and thus winter and spring cucurbit production in protected cultivation in the Great Lakes region (Papadopoulos and Gosselin 2007) could be a possible source of inoculum (Cohen et al., 2015; Ojiambo et al., 2017). Definitive evidence for this hypothesis has been elusive, though (Holmes et al., 2015).

The present data set and analysis do not provide evidence of an annually occurring, alternate source of *P. cubensis* in northern latitudes. In certain years downy mildew was observed in the Upper Midwest in June before the expected occurrence of a disease wavefront originating from southern Florida. However, two-source models with an observed or imputed alternate source placed in the Great Lakes region generally had the largest prediction errors and, in some cases, these errors were indeed so massively. We explored various spatial placements of alternate sources in northern latitudes (north, south, east, and west of Lake Erie) and timings for their appearance based on when sporangia of *P. cubensis* may be in the air (Granke and Hausbeck 2011; Granke et al., 2014). None of the observed or imputed sources led to appreciable improvements in prediction. Given the sparsity of the present data sets, absence of disease reports from greenhouses in the Cucurbit Downy Mildew ipmPIPE system, and our restriction of disease to the first planting of cucurbits, we caution that a lack of evidence for a second source in northern latitudes does not prove that one does not exist.

We did find support for an alternate inoculum source on the western edge of the Gulf of Mexico. This is perhaps unsurprising given that hosts of *P. cubensis* are present year-round in

frost-free areas along the southern Gulf Coast (Holmes et al., 2015). Depending on seasonal wind direction, the southwestern source is predicted to be a source for downy mildew in the southern plains, lower Midwestern states, and certain other regions of the southeastern U.S. Separate spatiotemporal analyses also suggest that inoculum sources in the southern U.S. outside of Florida may be responsible for disease outbreaks in more northern latitudes (Ojiambo and Holmes, 2011). In all years, predicted wavefronts from the southwestern source and Florida overlapped as early as May to June, potentially resulting in population admixture. Genetic evidence suggests that populations of *P. cubensis* in Florida may be differentiated from populations in Texas and certain other states (Quesada-Ocampo et al., 2012). A partial explanation for this genetic differentiation may be the presence of distinct overwintering populations and epidemic trajectories of downy mildew on the western and eastern edges of the Gulf of Mexico.

The multiple source modeling framework we present is most useful for posthoc analysis of epidemics rather than prediction. This is because parameter estimation requires an iterative procedure based on the known distribution of disease, which precludes prediction during an active epidemic. This has no bearing on one-source models, with or without anisotropy, which our analyses suggest may be adequate in some situations.

We introduced anisotropy in disease spread through the functions g_k that are estimated from prevailing wind directions at the epicenters. Wind direction and velocity at a primary inoculum source are associated with the shape of disease gradients when measured at the scale of individual or multiple fields (Rieux et al., 2014; Sackett and Mundt 2005). Wind speed and direction are also predictive of disease transmission of aerially dispersed pathogens at the mesoscale (Gent et al., 2019) and landscape-level (Sutrave et al., 2012). At these scales and the scales we evaluated, wind direction alone is only a simple correlate of a complex biophysical

process that may act along the entire path of dispersal (Allen-Sadler et al., 2019; Aylor 2003; Holmes et al., 2015; Ojiambo et al., 2015). More fundamentally, anisotropy in the early stages of an epidemic appears to be important for dispersal patterns that persist throughout the entire epidemic. It is unclear whether this observation is idiosyncratic to these particular data sets or suggestive of a more fundamental process of epidemic spread being heavily affected by properties of the initial disease epicenter.

We also point out that many other functions could be used to introduce anisotropy and van Putten et al. (2012) provide several useful statistical alternatives that do not explicitly consider wind. As we discussed above, physical process models could better capture anisotropy due to environmental factors but at the expense of greater data and computational requirements (Allen-Sader et al., 2019). Similarly, knowledge of host presence and their disease status in the landscape and more intensive placement and sampling of sentinel plots could enable one to develop time-varying anisotropy functions not possible here due to the extent of the cucurbit downy mildew data sets. Despite these limitations, the novelty and utility of our modeling framework are that it is tractable for estimating the generalized location and velocity of a disease front from sparsely sampled data with minimal data acquisition costs. Furthermore, when multiple sources exist the most probable source of the initial appearance of disease can be identified. These innovations make this modeling and estimation framework attractive for many problems central to dispersal, ecology of infectious disease, and management of epidemics.

References

1. Ojiambo PS, Gent DH, Mehra LK, Christie D, Magarey R. Focus expansion and stability of the spread parameter estimate of the power law model for dispersal gradients. PeerJ. 2017;5:e3465.
2. Clark JS, Lewis M, Horvath L. Invasion by extremes: population spread with variation in dispersal and reproduction. The American Naturalist. 2001;157(5):537–554.
3. Kot M, Lewis MA, van den Driessche P. Dispersal data and the spread of invading organisms. Ecology. 1996;77(7):2027–2042.
4. Ibrahim K, Nichols R, Hewitt G. Ibrahim KM, Nichols RA, Hewitt GM. Spatial patterns of genetic variation by different forms of dispersal during range expansion. Heredity. 1996;77:282–291.
5. Severns PM, Sackett KE, Farber DH, Mundt CC. Consequences of long-distance dispersal for epidemic spread: patterns, scaling, and mitigation. Plant Disease. 2019;103(2):177-191.
6. Wingen LU, Brown JKM, Shaw MW. The population genetic structure of clonal organisms generated by exponentially bounded and fat-tailed dispersal. Genetics. 2007;177(1):435-448.
7. Mundt CC, Sackett KE, Wallace LD, Cowger C, Dudley JP. Long-distance dispersal and accelerating waves of disease: empirical relationships. The American Naturalist. 2009;173(4):456-466.
8. Mundt CC, Sackett KE, Wallace LD, Cowger C, Dudley JP. Aerial dispersal and multiple scale spread of epidemic disease. EcoHealth. 2009;6(4):546-552.
9. Colucci SJ, Holmes GJ Downy mildew of cucurbits The Plant Health Instructor. PHI-I-2010-0825-01

10. Jaing C, Thissen J, Morrison M, Dillon MB, Waters SM, Graham GT, et al. Sierra Nevada sweep: metagenomic measurements of bioaerosols vertically distributed across the troposphere. Scientific Reports. 2020;10(1).
11. Ojiambo PS, Holmes GJ, Britton W, Babadoost M, Bost SC, Boyles R, et al. Cucurbit Downy Mildew ipmPIPE: a next generation web-based interactive tool for disease management and extension outreach. Plant Health Progress. 2011;12(1):26.
12. Ojiambo PS, Holmes GJ. Spatiotemporal spread of cucurbit downy mildew in the eastern United States. Phytopathology. 2011;101(4):451-461.
13. Ojiambo PS, Gent DH, Quesada-Ocampo LM, Hausbeck MK, Holmes GJ. Epidemiology and population biology of *Pseudoperonospora cubensis*: a model system for management of downy mildews. Annual Review of Phytopathology. 2015;53(1):223-246.
14. Cohen Y, den Langenberg KMV, Wehner TC, Ojiambo PS, Hausbeck M, Quesada-Ocampo LM, et al. Resurgence of *Pseudoperonospora cubensis*: the causal agent of cucurbit downy mildew. Phytopathology. 2015;105(7):998-1012.
15. Madden LV, Hughes G, van den Bosch F. The Study of Plant Disease Epidemics. The American Phytopathological Society; 2007.
16. Gregory PH. Interpreting plant disease dispersal gradients. Annual Review of Phytopathology. 1968;6(1):189-212.
17. Holmes GJ, Ojiambo PS, Hausbeck MK, Quesada-Ocampo L, Keinath AP. Resurgence of cucurbit downy mildew in the United States: a watershed event for research and extension. Plant Disease. 2015;99(4):428-441.
18. Taylor PD, Fahrig L, Henein K, Merriam GW. Connectivity is a vital element of landscape structure. Oikos. 1993;68:571-573.

19. Margosian ML, Garrett KA, Hutchinson JMS, With KA. Connectivity of the American agricultural landscape: assessing the national risk of crop pest and disease spread. BioScience. 2009;59(2):141-151.
20. van Putten B, Visser MD, Muller-Landau HC, Jansen PA. Distorted-distance models for directional dispersal: a general framework with application to a wind-dispersed tree. Methods in Ecology and Evolution. 2012;3(4):642-652.
21. Soubeyrand S, Enjalbert J, Sanchez A, Sache I. Anisotropy, in density and in distance, of the dispersal of yellow rust of wheat: experiments in large field plots and estimation. Phytopathology. 2007;97(10):1315-1324.
22. Rieux A, Soubeyrand S, Bonnot F, Klein EK, Ngando JE, Mehl A, et al. Long-distance wind-dispersal of spores in a fungal plant pathogen: estimation of anisotropic dispersal kernels from an extensive field experiment. PLoS One. 2014;9(8):e103225.
23. Cowger C, Wallace LD, Mundt CC. Velocity of spread of wheat stripe rust epidemics. Phytopathology. 2005;95(9):972–982.
24. Waggoner PE. Weather, space, time, and chance of infection. Phytopathology. 1962;52:1100-1108.
25. Meyer M, Burgin L, Hort MC, Hodson DP, Gilligan CA. Large-scale atmospheric dispersal simulations identify likely airborne incursion routes of wheat stem rust into Ethiopia. Phytopathology. 2017;107(10):1175-1186.
26. Savory EA, Granke LL, Quesada-Ocampo LM, Varbanova M, Hausbeck MK, Day B. The cucurbit downy mildew pathogen *Pseudoperonospora cubensis*. Molecular Plant Pathology. 2010;12(3):217-226.

27. Thomas A, Carbone I, Cohen Y, Ojiambo PS. Occurrence and distribution of mating types of *Pseudoperonospora cubensis* in the United States.Phytopathology. 2017;107(3):313-321.
28. Smith A, Lott N, Vose R. The Integrated Surface Database: recent developments and partnerships. Bulletin of the American Meteorological Society. 2011;92(6):704-708.
29. Meentemeyer RK, Haas SE, Vaclavık T. Landscape epidemiology of emerging infectious diseases in natural and human-altered ecosystems. Annual Review of Phytopathology. 2012;50(1):379-402.
30. Xing Y, Nopsa JFH, Andersen KF, Andrade-Piedra JL, Beed FD, Blomme G, et al. Global cropland connectivity: a risk factor for invasion and saturation by emerging pathogens and pests. BioScience. 2020.
31. Farber DH, Medlock J, Mundt CC. Local dispersal of *Puccinia striiformis* f. sp. tritici from isolated source lesions. Plant Pathology. 2016;66(1):28-37.
32. Mundt CC, Sackett KE. Spatial scaling relationships for spread of disease caused by a wind-dispersed plant pathogen. Ecosphere. 2012;3(3):art24.
33. Gent DH, Bhattacharyya S, Ruiz T. Prediction of spread and regional development of hop powdery mildew: a network analysis. Phytopathology. 2019;109(8):1392-1403.
34. Sutrave S, Scoglio C, Isard SA, Hutchinson JMS, Garrett KA. Identifying highly connected counties compensates for resource limitations when evaluating national spread of an invasive pathogen. PLoS ONE. 2012;7(6):e37793.
35. Gent DH, Mahaffee WF, McRoberts N, Pfender WF. The use and role of predictive systems in disease management. Annual Review of Phytopathology. 2013;51(1):267-289.

36. Mundt CC, Wallace LD, Allen TW, Hollier CA, Kemerait RC, Sikora E. Initial epidemic area is strongly associated with the yearly extent of soybean rust spread in North America. Biological Invasions. 2013;15(7):1431-1438.
37. Filho AB, Inoue-Nagata AK, Bassanezi RB, Belasque J, Amorim L, Macedo MA, et al. The importance of primary inoculum and area-wide disease management to crop health and food security. Food Security. 2016;8(1):221-238.
38. Gent DH, Mahaffee WF, Turechek WW, Ocamb CM, Twomey MC, Woods JL, et al. Risk factors for bud perennation of *Podosphaera macularis* on hop. Phytopathology. 2019;109(1):74-83.
39. Graham J, Gottwald T, Setamou M. Status of Huanglongbing (HLB) outbreaks in Florida, California and Texas. Tropical Plant Pathology. 2020;45(3):265-278.
40. Nathan R, Perry G, Cronin JT, Strand AE, Cain ML. Methods for estimating long-distance dispersal. Oikos. 2003;103(2):261-273.
41. Irwin M. Implications of movement in developing and deploying integrated pest management strategies. Agricultural and Forest Meteorology. 1999;97(4):235-248.
42. Naegele RP, Quesada-Ocampo LM, Kurjan JD, Saude C, Hausbeck MK. Regional and temporal population structure of *Pseudoperonospora cubensis* in Michigan and Ontario. Phytopathology. 2016;106(4):372-379.
43. Papadopoulos T, Gosselin A. Greenhouse vegetable production in Canada. Chronica Horticulture. 2007;47(3):23-28.
44. Granke LL, Hausbeck MK. Dynamics of *Pseudoperonospora cubensis* sporangia in commercial cucurbit fields in Michigan. Plant Disease. 2011;95(11):1392-1400.

45. Granke LL, Morrice JJ, Hausbeck MK. Relationships between airborne *Pseudoperonospora cubensis* sporangia, environmental conditions, and cucumber downy mildew severity. Plant Disease. 2014;98(5):674-681.
46. Quesada-Ocampo LM, Granke LL, Olsen J, Gutting HC, Runge F, Thines M, et al. The genetic structure of Pseudoperonospora cubensis populations. Plant Disease. 2012;96(10):1459-1470.
47. Sackett KE, Mundt CC. Primary disease gradients of wheat stripe rust in large field plots. Phytopathology. 2005;95(9):983-991.
48. Allen-Sader C, Thurston W, Meyer M, Nure E, Bacha N, Alemayehu Y, et al. An early warning system to predict and mitigate wheat rust diseases in Ethiopia. Environmental Research Letters. 2019;14(11):115004.
49. Aylor DE. Spread of plant disease on a continental scale: role of aerial dispersal of pathogens. Ecology. 2003;84(8):1989-1997.

Tables

Table 4.1. Notations used in the multi-source model. The horizontal line divides notations for data quantities (above) from notations for model quantities (below). In the text, subscripts i are appended to the data notations to indicate the corresponding quantity for the ith observation. Similarly, hats are placed over the model quantities to indicate estimates (e.g., $\hat{c}_k$).

Notation	Description
$x^{(k)}$	kth source location (Cartesian)
$r^{(k)}$	distance to kth source (km)
$\phi^{(k)}$	angle to kth source (radians)
t	time (day of year)
K	number of sources
g_k	kth directional anisotropy function
$p^{(k)}$	probability that disease is caused by kth source
c_k	kth regression intercept
$-M_k$	kth regression parameter for time
$h_k(\cdot)$	normalizing function for kth regression model
$\epsilon^{(k)}$	error term in kth regression model
σ_k^2	error variance in kth regression model
$\sum_b \beta_b^{(k)} s_b(\cdot)$	basis function approximation for h_k

Table 4.2. The mean parameter estimates and standard deviation for two-source models fit to simulated data. Two sets of estimates are reported corresponding to a (0,0) placement for a first source and a (2000,2000) placement for a second source.

			Source 1				Source 2			
Start time	n	σ^2	Intercept	Time	Basis 1	Basis 2	Intercept	Time	Basis 1	Basis 2
True values			4.85	0.03	0.5	0	3.75	0.02	0	-1
Synchronous	100	0.5	4.849	0.030	0.498	-0.003	3.768	0.020	-0.005	-0.968
			(0.322)	(0.003)	(0.178)	(0.175)	(0.556)	(0.005)	(0.295)	(0.313)
		1	5.100	0.028	0.489	-0.067	4.435	0.017	-0.011	-0.79
			(0.777)	(0.007)	(0.383)	(0.366)	(6.867)	(0.015)	(1.003)	(6.579)
	500	0.5	4.860	0.030	0.497	0.001	3.783	0.020	-0.001	-0.987
			(0.142)	(0.001)	(0.077)	(0.075)	(0.225)	(0.002)	(0.123)	(0.126)
		1	5.561	0.024	0.501	-0.189	4.032	0.016	0.005	-0.700
			(0.373)	(0.003)	(0.158)	(0.155)	(0.709)	(0.007)	(0.376)	(0.456)
Asynchronous	100	0.5	4.860	0.030	0.505	-0.005	4.015	0.018	0.005	-0.949
			(0.334)	(0.003)	(0.177)	(0.179)	(1.655)	(0.013)	(0.343)	(0.350)
		1	5.117	0.028	0.514	-0.006	6.378	0.001	-0.039	-0.317
			(0.791)	(0.007)	(0.441)	(0.390)	(6.858)	(0.026)	(1.214)	(6.007)
	500	0.5	4.935	0.029	0.498	0.003	3.958	0.018	-0.001	-0.931
			(0.166)	(0.002)	(0.075)	(0.074)	(0.712)	(0.006)	(0.158)	(0.167)
		1	5.180	0.027	0.502	0.003	8.563	-0.011	-0.014	-0.808
			(0.299)	(0.003)	(0.161)	(0.145)	(18.319)	(0.017)	(0.742)	(18.197)

Table 4.3. Parameter estimates and 95% confidence intervals for one- and two-source models fit to 2008 data (n = 25). In the two-source models, two sets of estimates are reported corresponding to a northern placement and a southwestern placement for the alternate (non-FL) source. No basis parameters are reported for the isotropic models, since these terms are only included in the anisotropic models.

Model (Source)	Intercept ($\hat{c}_k$)	Time ($-\widehat{M}_k$)	Basis 1 ($\hat{\beta}_1^{(k)}$)	Basis 2 ($\hat{\beta}_2^{(k)}$)
IOS (FL)	3.052	0.020		
	(2.451, 3.654)	(0.017, 0.024)	---	---
ITS-SW (FL)	2.943	0.021		
	(2.464, 3.421)	(0.018, 0.024)	---	---
ITS-SW (TX)	1.784	0.024		
	(0.583, 2.986)	(0.017, 0.031)	---	---
ITS-N (FL)	2.975	0.021		
	(2.370, 3.581)	(0.017, 0.025)	---	---
ITS-N (OH)	24.675	-0.089		
	(-15.374, 64.725)	(-0.281, 0.103)	---	---
AOS (FL)	4.921	0.021	-0.234	0.207
	(4.273, 5.568)	(0.017, 0.026)	(-0.679, 0.212)	(-0.296, 0.710)
ATS-SW (FL)	4.887	0.021	-0.189	0.297
	(4.287, 5.488)	(0.016, 0.026)	(-0.572, 0.194)	(-0.293, 0.886)
ATS-SW (TX)	-18.597	0.033	6.894	18.280
	(-22.652, -14.543)	(0.033, 0.033)	(3.001, 10.787)	(18.023, 18.537)
ATS-N (FL)	4.885	0.022	-0.194	0.191
	(4.241, 5.528)	(0.017, 0.026)	(-0.608, 0.219)	(-0.264, 0.647)
ATS-N (OH)	7.780	-0.009	5.089	-0.475
	(-117.549, 133.110)	(-0.562, 0.543)	(-19.883, 30.062)	(-10.783, 9.833)

Table 4.4. Parameter estimates and 95% confidence intervals for one- and two-source models fit to 2009 data (n = 65). In the two-source models, two sets of estimates are reported corresponding to a northern placement and a southwestern placement for the alternate (non-FL) source. No basis parameters are reported for the isotropic models, since these terms are only included in the anisotropic models.

Model (Source)	Intercept ($\hat{c}_k$)	Time ($-\widehat{M}_k$)	Basis 1 ($\hat{\beta}_1^{(k)}$)	Basis 2 ($\hat{\beta}_2^{(k)}$)
IOS (FL)	4.524	0.014		
	(3.744, 5.304)	(0.010, 0.018)	---	---
ITS-SW (FL)	4.194	0.016		
	(3.516, 4.873)	(0.012, 0.019)	---	---
ITS-SW (TX)	5.092	0.009		
	(4.153, 6.032)	(0.004, 0.015)	---	---
ITS-N (FL)	2.921	0.023		
	(2.466, 3.377)	(0.020, 0.026)	---	---
ITS-N (OH)	7.795	-0.007		
	(5.453, 10.137)	(-0.020, 0.005)	---	---
AOS (FL)	8.531	0.014	2.555	-2.334
	(7.262, 9.799)	(0.010, 0.019)	(1.153, 3.958)	(-3.758, -0.910)
ATS-SW (FL)	8.179	0.014	2.028	-1.636
	(7.011, 9.347)	(0.010, 0.018)	(0.736, 3.319)	(-2.968, -0.304)
ATS-SW (TX)	5.609	0.027	5.328	-3.062
	(4.515, 6.704)	(0.017, 0.036)	(4.603, 6.054)	(-4.714, -1.409)
ATS-N (FL)	4.873	0.021	-0.599	1.424
	(4.020, 5.726)	(0.017, 0.024)	(-1.434, 0.236)	(0.556, 2.293)
ATS-N (OH)	9.148	-0.005	0.308	-0.396
	(7.277, 11.019)	(-0.015, 0.004)	(-0.089, 0.706)	(-0.818, 0.026)

Table 4.5. Parameter estimates and 95% confidence intervals for one- and two-source models fit to 2010 data (n = 28). In the two-source models, two sets of estimates are reported corresponding to a northern placement and a southwestern placement for the alternate (non-FL) source. No basis parameters are reported for the isotropic models, since these terms are only included in the anisotropic models. In this year, no alternate source model is estimated for the southwest location, as nearly all data points are attributed to the FL source during model fitting.

Model (Source)	Intercept ($\hat{c}_k$)	Time ($-\widehat{M}_k$)	Basis 1 ($\hat{\beta}_1^{(k)}$)	Basis 2 ($\hat{\beta}_2^{(k)}$)
IOS (FL)	3.433	0.017		
	(2.200, 4.665)	(0.011, 0.024)	---	---
ITS-SW (FL)	3.564	0.016		
	(2.381, 4.747)	(0.010, 0.023)	---	---
ITS-N (FL)	4.337	0.011		
	(3.342, 5.331)	(0.005, 0.017)	---	---
ITS-N (NY)	5.363	0.004		
	(1.612, 9.115)	(-0.015, 0.023)	---	---
AOS (FL)	4.967	0.020	0.226	-0.129
	(3.116, 6.817)	(0.009, 0.032)	(-0.475, 0.927)	(-0.814, 0.557)
ATS-SW (FL)	5.218	0.019	0.292	-0.053
	(3.384, 7.051)	(0.007, 0.030)	(-0.388, 0.972)	(-0.743, 0.636)
ATS-N (FL)	5.745	0.015	0.364	-0.268
	(4.536, 6.955)	(0.007, 0.022)	(-0.081, 0.809)	(-0.680, 0.144)
ATS-N (NY)	10.675	-0.014	-0.040	0.559
	(5.278, 16.073)	(-0.040, 0.012)	(-0.509, 0.429)	(-0.565, 1.684)

Table 4.6. Root mean square errors of time and distance for isotropic and anisotropic one-source models. The source location and time of appearance, shown in the table, are the earliest occurrences among all sentinel plots in the data for the corresponding year.

	Florida source		Isotropic		Anisotropic	
Year (*n*)	County, State	Date	Time (days)	Distance (km)	Time (days)	Distance (km)
2008 (25)	Collier, FL	02-18	12.38	271.73	11.62	273.61
2009 (65)	Miami-Dade, FL	03-23	26.14	443.78	24.79	411.62
2010 (28)	Alachua, FL	03-24	24.44	325.09	23.50	392.40

Table 4.7. Root mean square errors of time and distance for isotropic and anisotropic two-source models for several alternate source locations. The other of the two sources is placed in Florida in the same location and time as in the one-source model.

	Alternate (non-FL) source		Isotropic		Anisotropic	
Year (*n*)	County, State	Date	Time (days)	Distance (km)	Time (days)	Distance (km)
2008 (25)	Vernon, LA	06-12	9.61	318.11	12.42	170.04
	Brazos, TX	05-06	9.88	238.26	9.57	237.38
	Hidalgo, TX	05-06	9.61	236.70	10.36	240.88
	Sandusky, OH	07-20	11.20	244.74	30.42	267.16
	Huron, OH	06-03	45.95	300.67	61.31	280.83
	Niagara, NY	06-03	17.68	229.13	124.26	153.92
2009 (65)	Payne, OK	06-16	26.23	379.49	21.14	356.59
	Brazos, TX	05-07	47.29	452.58	35.08	669.44
	Hidalgo, TX	05-07	23.48	361.31	23.12	360.34
	Huron, OH	06-05	81.24	365.92	88.36	344.33
	Sandusky, OH	06-04	1676.48	390.29	151.06	370.96
	Niagara, NY	06-04	20517.83	411.95	28.40	245.95
2010 (28)	Brazos, TX	07-12	25.01	300.38	25.27	374.55
	Brazos, TX	05-07	24.54	294.76	24.79	367.55
	Hidalgo, TX	05-07	24.37	315.55	24.84	391.08
	Wayne, OH	07-03	57.03	251.97	109.26	314.68
	Sandusky, OH	06-04	2492.48	270.42	61.12	295.33
	Niagara, NY	06-04	52.75	200.48	24.03	192.53

Figures

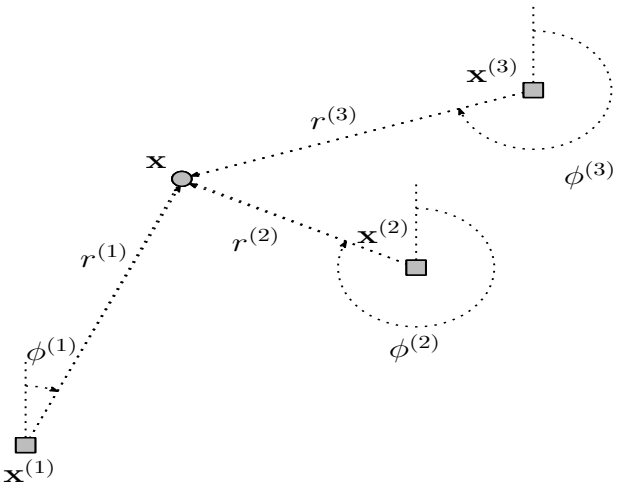


Figure 4.1. Depiction of data representation. A single location x is shown relative to three source points $x^{(1)}, x^{(2)}$, and $x^{(3)}$, and the polar coordinates $\left(r^{(1)}, \phi^{(1)}\right), \left(r^{(2)}, \phi^{(2)}\right)$, and $\left(r^{(3)}, \phi^{(3)}\right)$ label the distances and angles to each source.

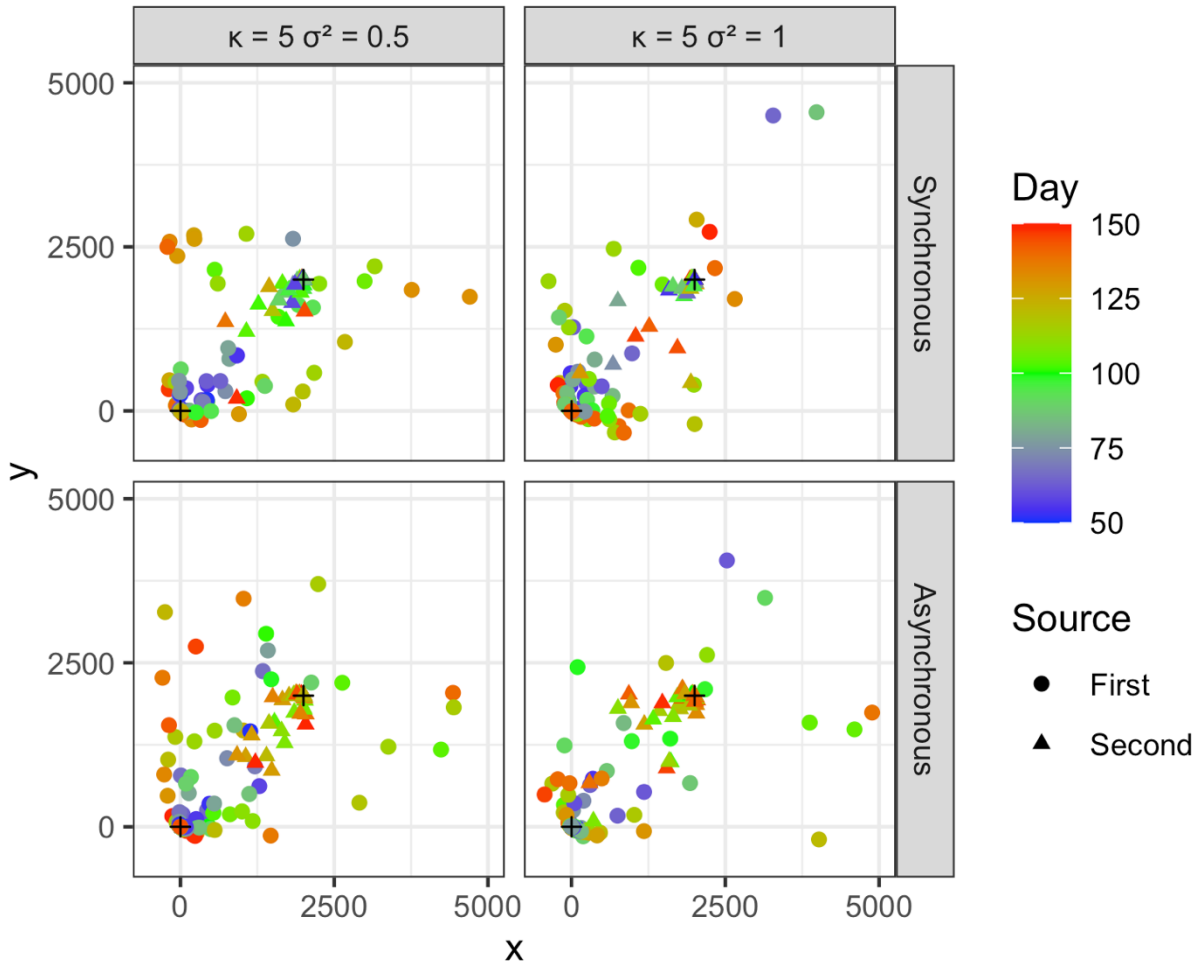


Figure 4.2. **The locations of observations from one simulation for $\kappa = 5$, $\sigma^2 = 0.5$ and 1.** The results are shown for temporally synchronous and asynchronous epidemics. The x and y-axis scales are set to -500 and 5000 for better visualization.

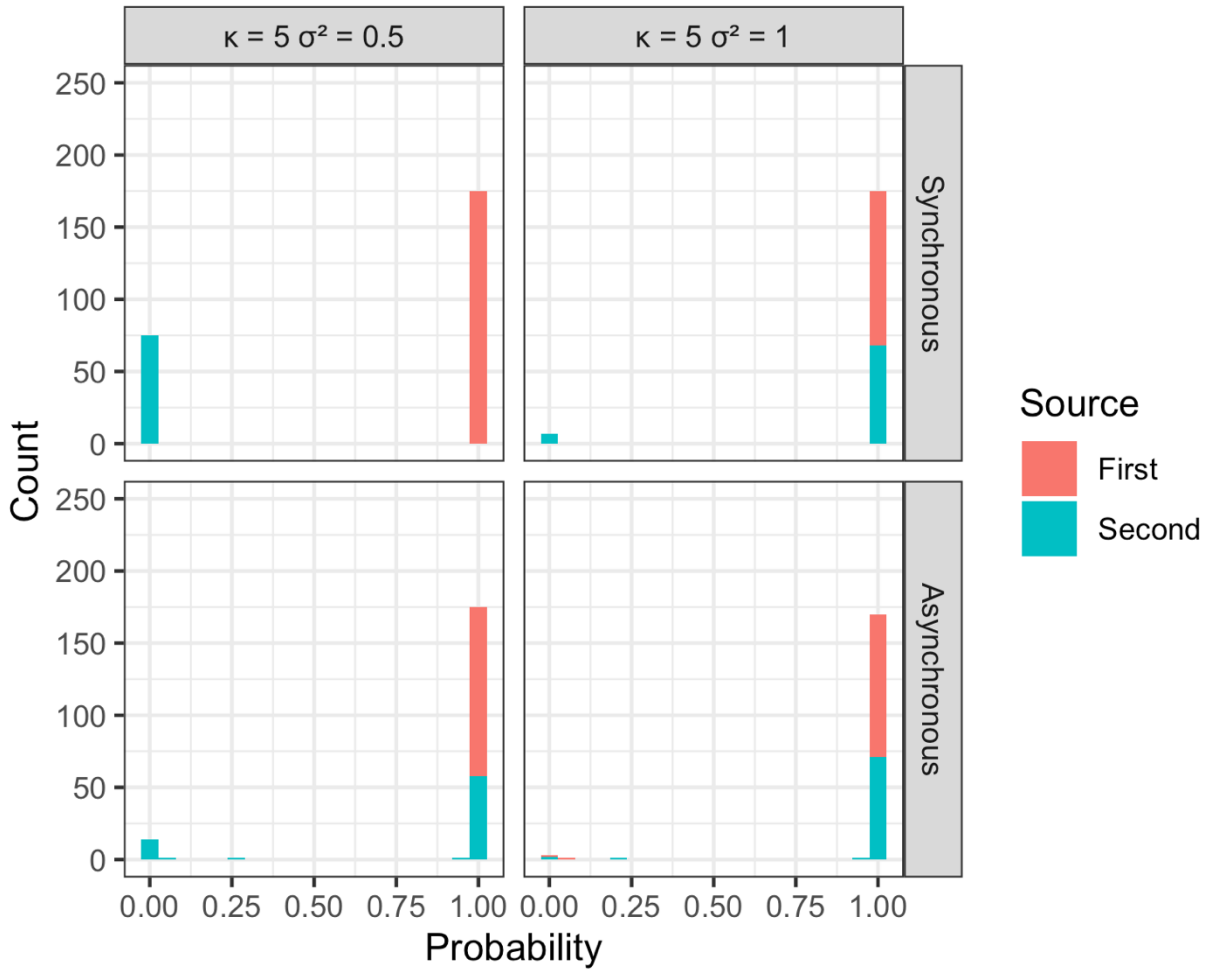


Figure 4.3. **Estimated probability of disease source for each of the n = 250 observations in individual, representative simulations with κ = 5, σ^2 = 0.5 and 1.** Results are shown for temporally synchronous and asynchronous epidemics.

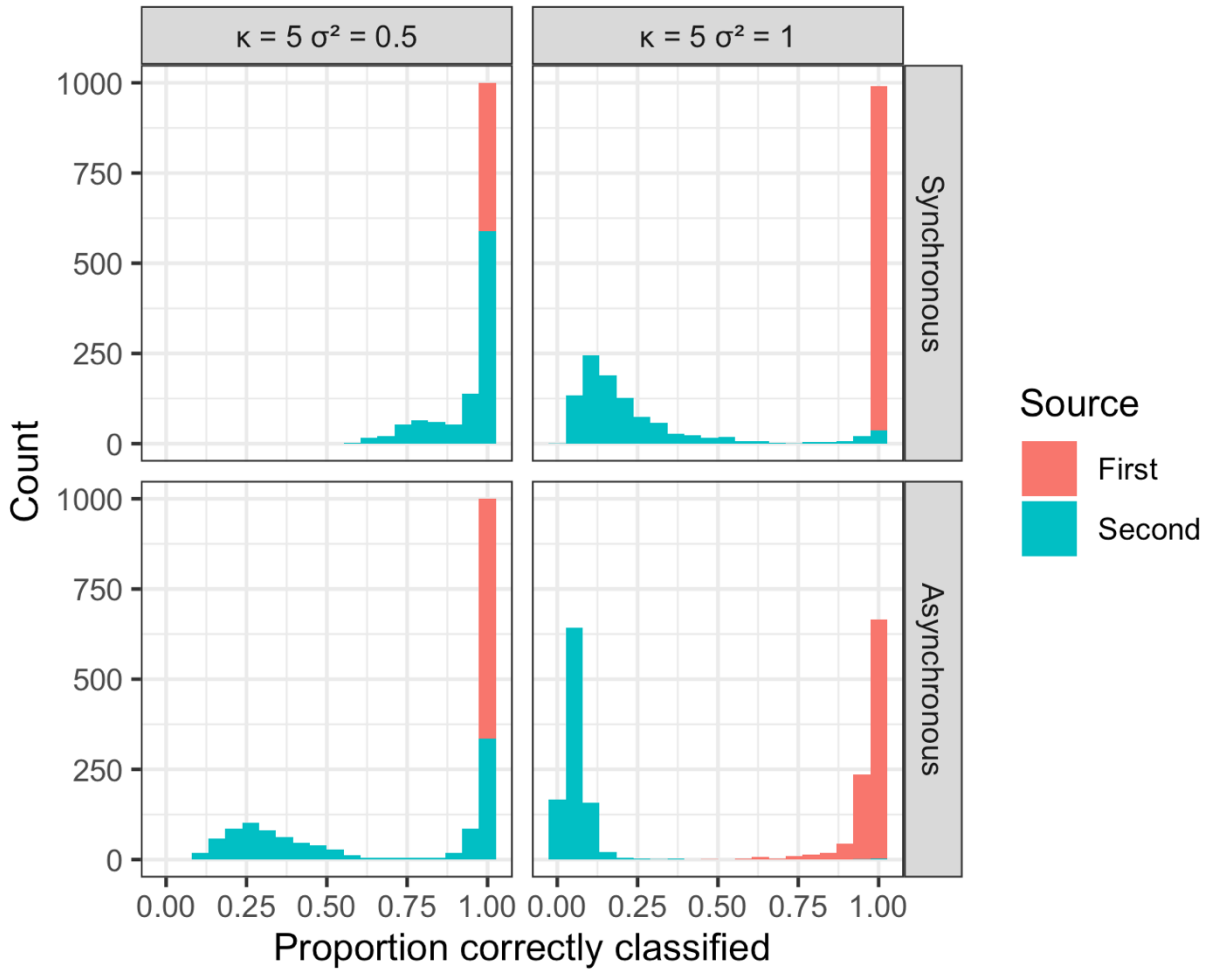


Figure 4.4. **Results from a simulation experiment with n = 250; κ = 5, two levels of error variance (σ^2), and temporally synchronous or asynchronous epidemics.** The histogram summarizes the mean proportion of disease correctly assigned to the true source over 1000 simulations.

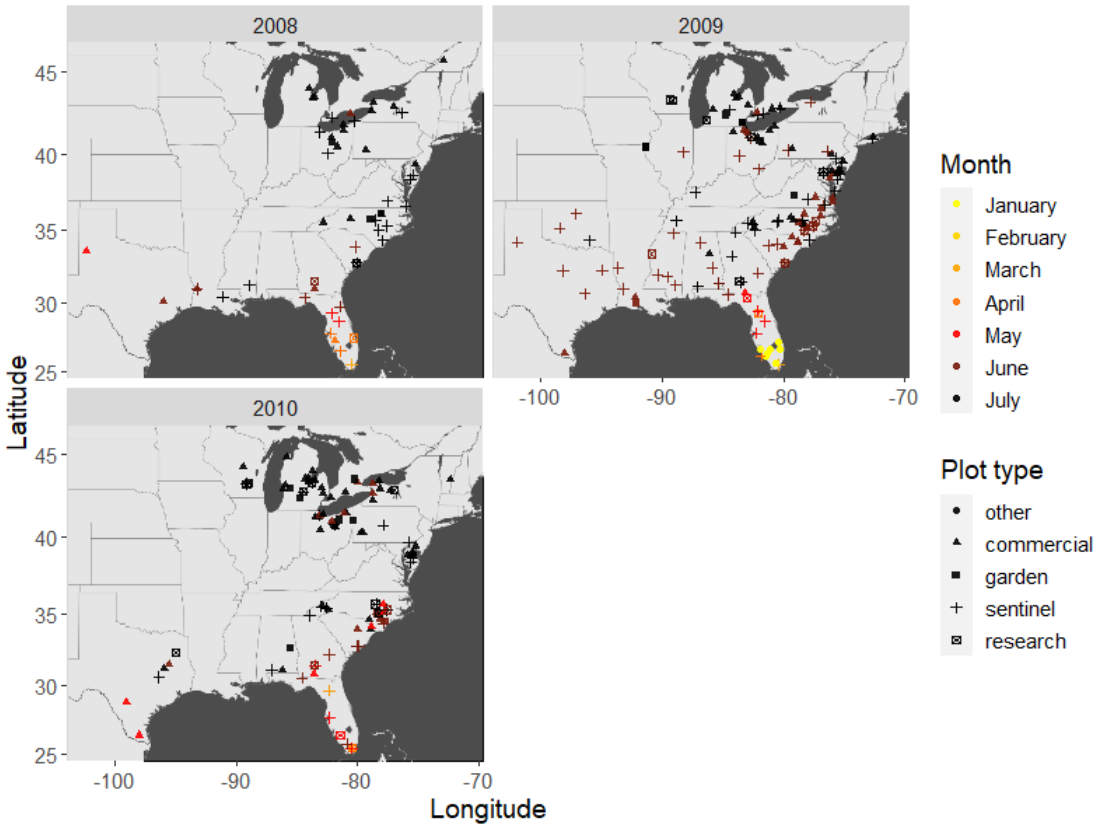


Figure 4.5. Disease reports from 2008 to 2010 plotted by location, reported symptom date, and plot type.

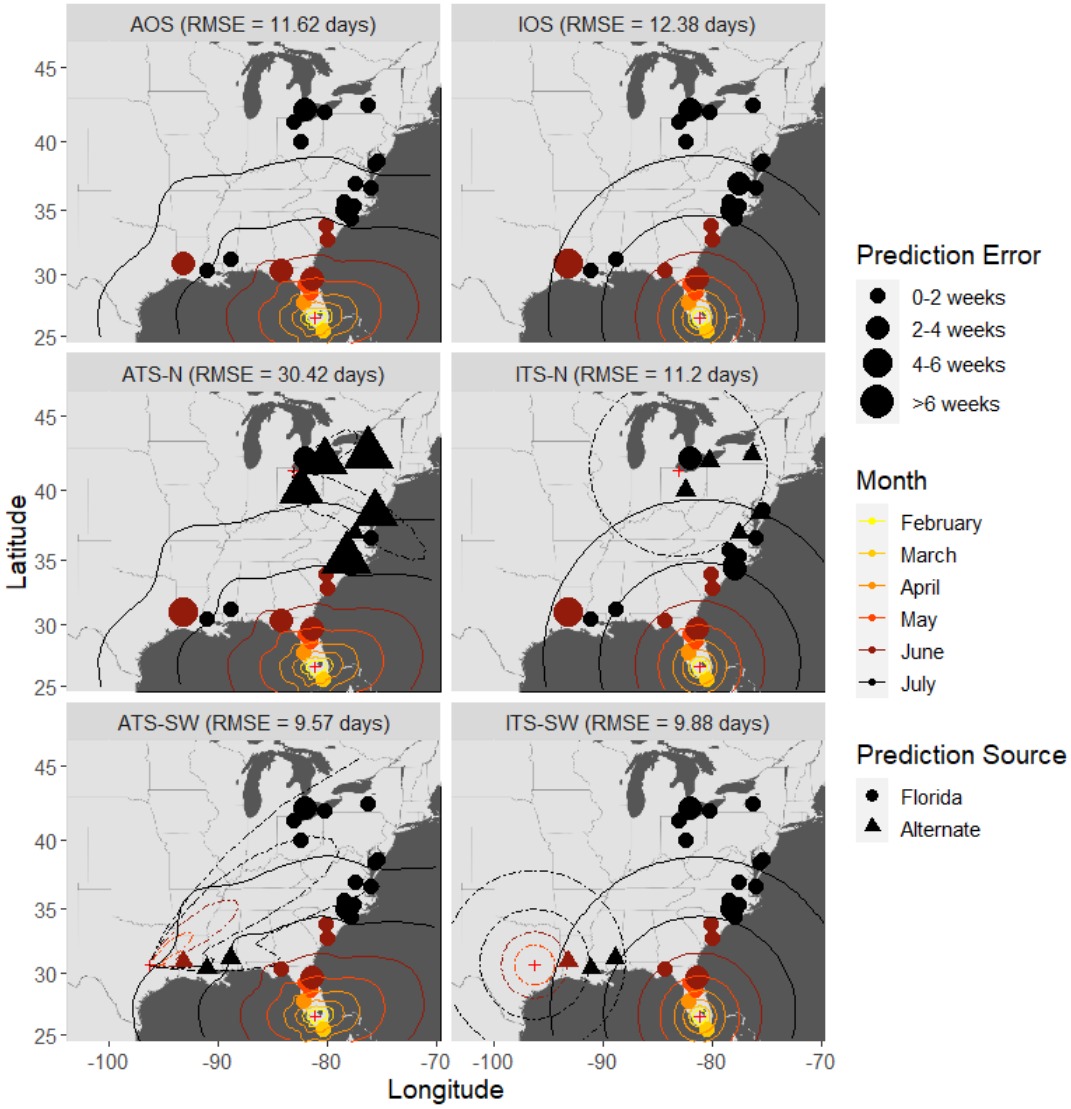


Figure 4.6. Time-of-occurrence prediction errors for predictions from isotropic and anisotropic one- and two-source models fit to data from sentinel plots in 2008: with contours representing estimated disease front over time according to the models. The two-source models were each fit with two alternate (non-FL) source locations: a northern and a southwestern location. Each panel shows results according to a different model: isotropic one-source (IOS); isotropic two-source (ITS); anisotropic one-source (AOS); and anisotropic two-source (ATS).

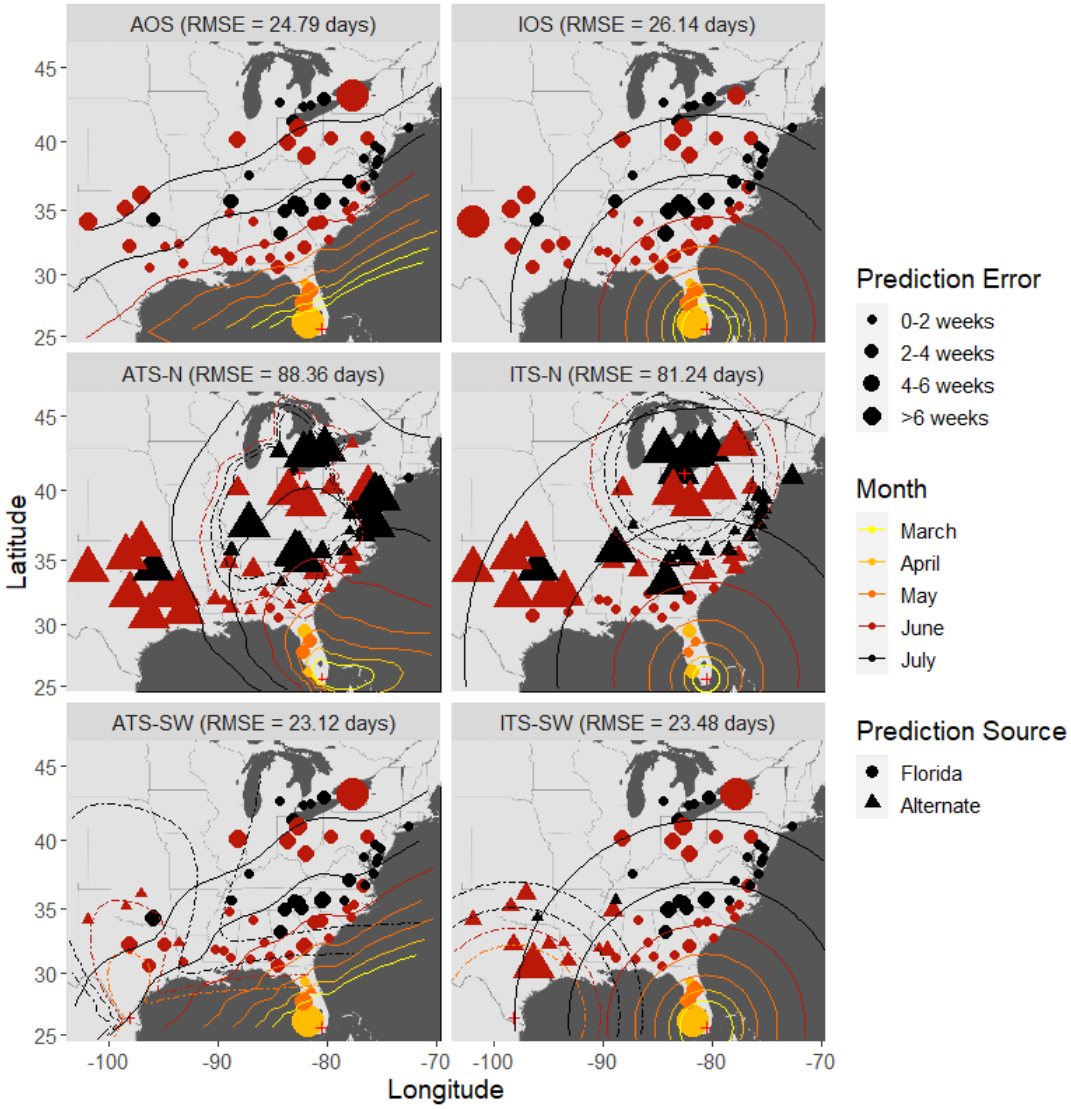


Figure 4.7. Time-of-occurrence prediction errors for predictions from isotropic and anisotropic one- and two-source models fit to data from sentinel plots in 2009: with contours representing estimated disease front over time according to the models. The two-source models were each fit with two alternate (non-FL) source locations: a northern and a southwestern location. Each panel shows results according to a different model: isotropic one-source (IOS); isotropic two-source (ITS); anisotropic one-source (AOS); and anisotropic two-source (ATS).

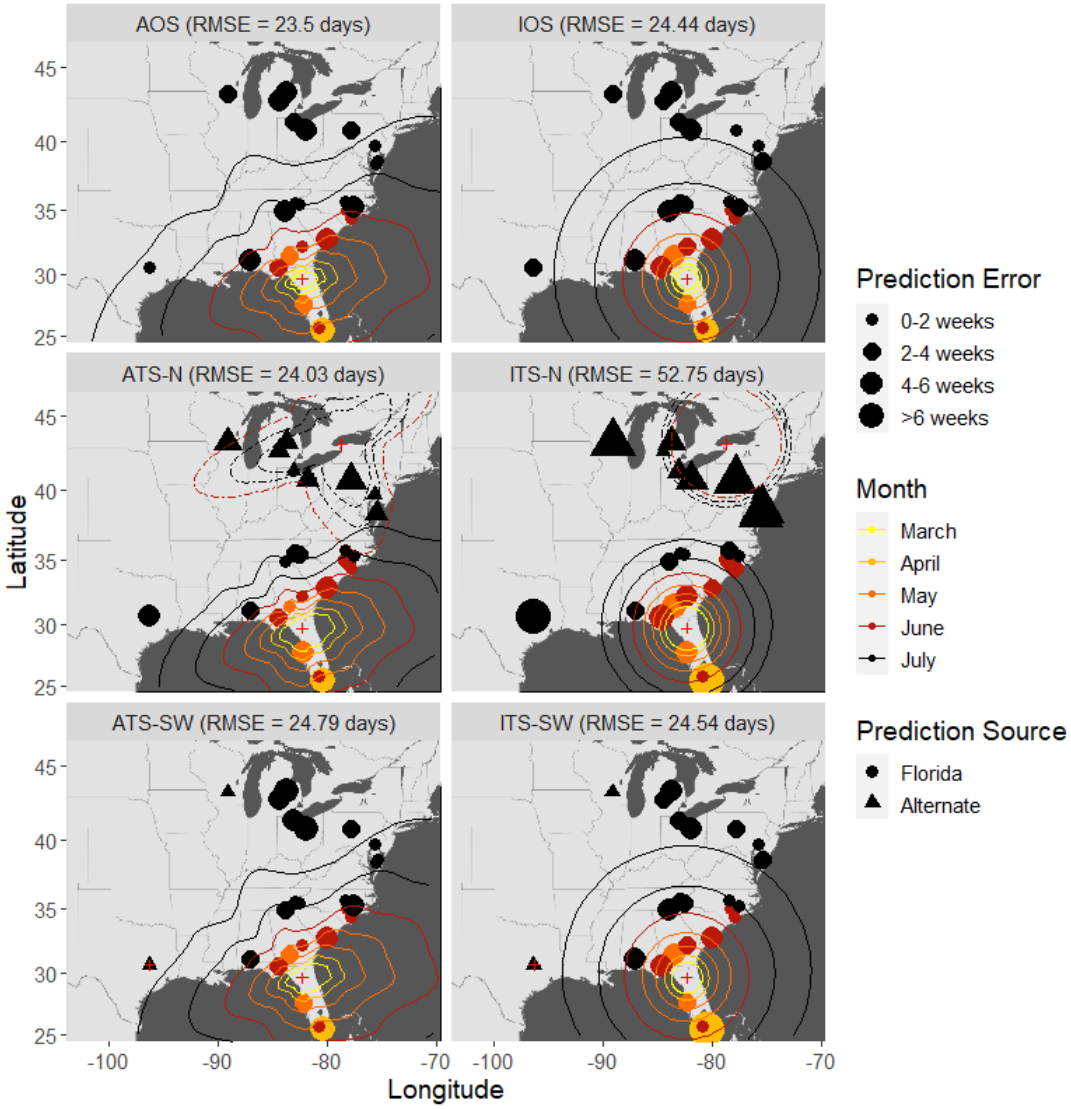


Figure 4.8. Time-of-occurrence prediction errors for predictions from isotropic and anisotropic one- and two-source models fit to data from sentinel plots in 2010: with contours representing estimated disease front over time according to the models. The two-source models were each fit with two alternate (non-FL) source locations: a northern and a southwestern location. Each panel shows results according to a different model: isotropic one-source (IOS); isotropic two-source (ITS); anisotropic one-source (AOS); and anisotropic two-source (ATS).

Supplemental Tables

Table S4.1. The mean parameter estimates and standard deviation for two-source models fit to simulated data. Two sets of estimates are reported corresponding to a (0,0) placement for a first source and a (2000,2000) placement for a second source

			Source 1				Source 2			
Start time	n	σ^2	Intercept	Time	Basis 1	Basis 2	Intercept	Time	Basis 1	Basis 2
True values			4.85	0.03	0.5	0	3.75	0.02	0	-1
Synchronous	100	0.5	4.862	0.03	0.507	-0.007	3.797	0.019	0.011	-0.985
			(0.216)	(0.002)	(0.13)	(0.124)	(0.348)	(0.003)	(0.2)	(0.202)
		1	5.291	0.026	0.493	-0.119	4.009	0.018	0.026	-0.596
			(0.585)	(0.005)	(0.252)	(0.253)	(1.193)	(0.011)	(0.626)	(0.914)
	500	0.5	4.869	0.03	0.504	-0.006	3.787	0.019	0.001	-0.986
			(0.103)	(0.001)	(0.054)	(0.055)	(0.155)	(0.002)	(0.096)	(0.096)
		1	5.691	0.023	0.499	-0.219	3.985	0.017	0.011	-0.699
			(0.204)	(0.002)	(0.118)	(0.104)	(0.523)	(0.005)	(0.277)	(0.33)
Asynchronous	100	0.5	4.889	0.03	0.498	0.003	3.943	0.018	0.002	-0.949
			(0.221)	(0.002)	(0.123)	(0.121)	(1.052)	(0.008)	(0.221)	(0.245)
		1	5.227	0.027	0.498	0.001	6.771	-0.005	-0.006	0.167
			(0.516)	(0.004)	(0.263)	(0.233)	(13.537)	(0.021)	(0.934)	(13.342)
	500	0.5	4.987	0.029	0.501	0.004	4.027	0.017	-0.002	-0.931
			(0.127)	(0.001)	(0.054)	(0.050)	(0.55)	(0.004)	(0.126)	(0.126)
		1	5.201	0.027	0.496	0.001	9.534	-0.018	-0.010	-0.888
			(0.207)	(0.002)	(0.109)	(0.094)	(12.000)	(0.014)	(0.485)	(11.952)

Supplemental Figures

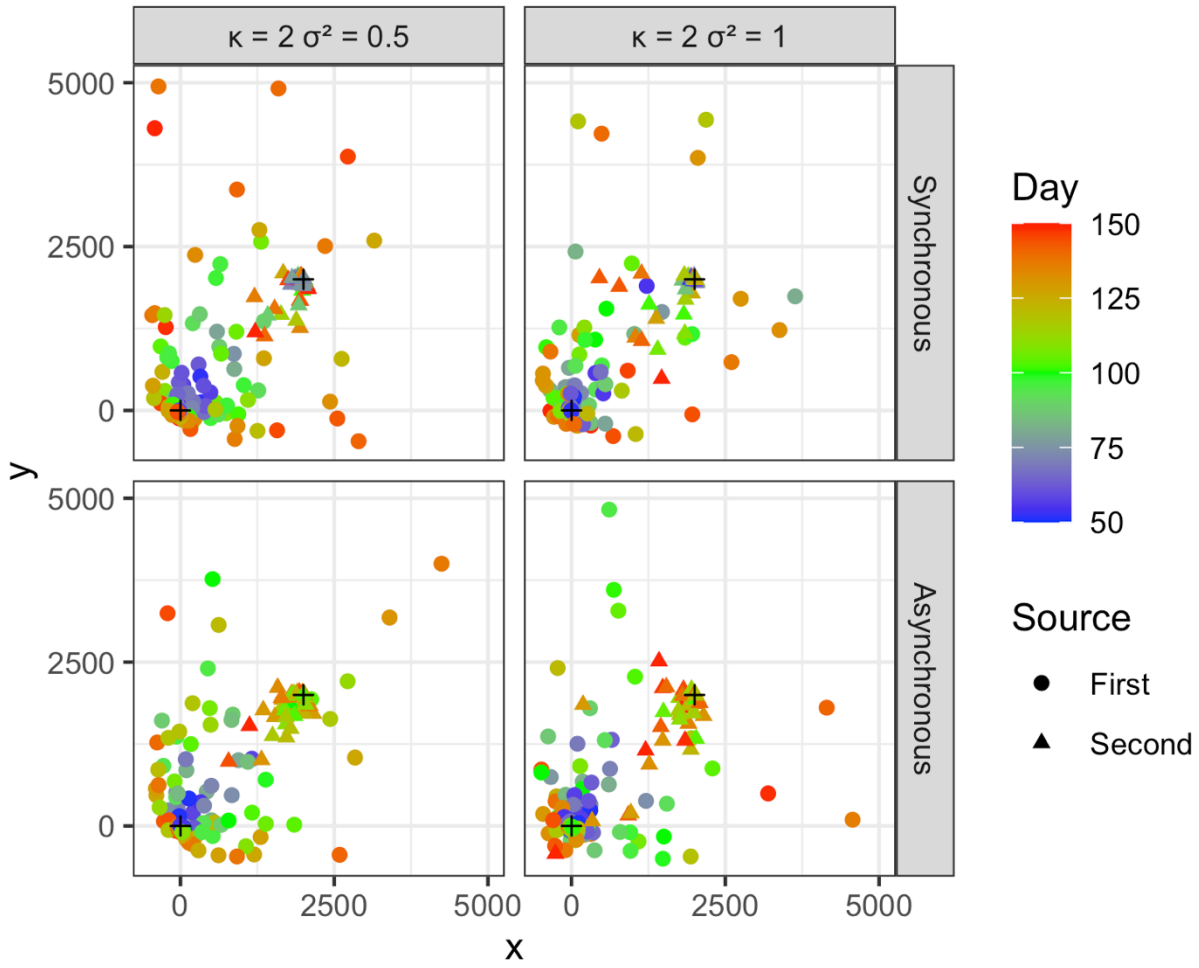


Figure S4.1. The locations of observations from one simulation for $\kappa = 2$, $\sigma^2 = 0.5$ and 1. The results are shown for synchronous and asynchronous start times. The plus signs, circles, and triangles represent the sources, the locations attributed to the first source, and the locations attributed to the second source respectively. The points are color-coded based on the day of the year. The x and y graph limits are set to -500 and 5000 for better visualization.

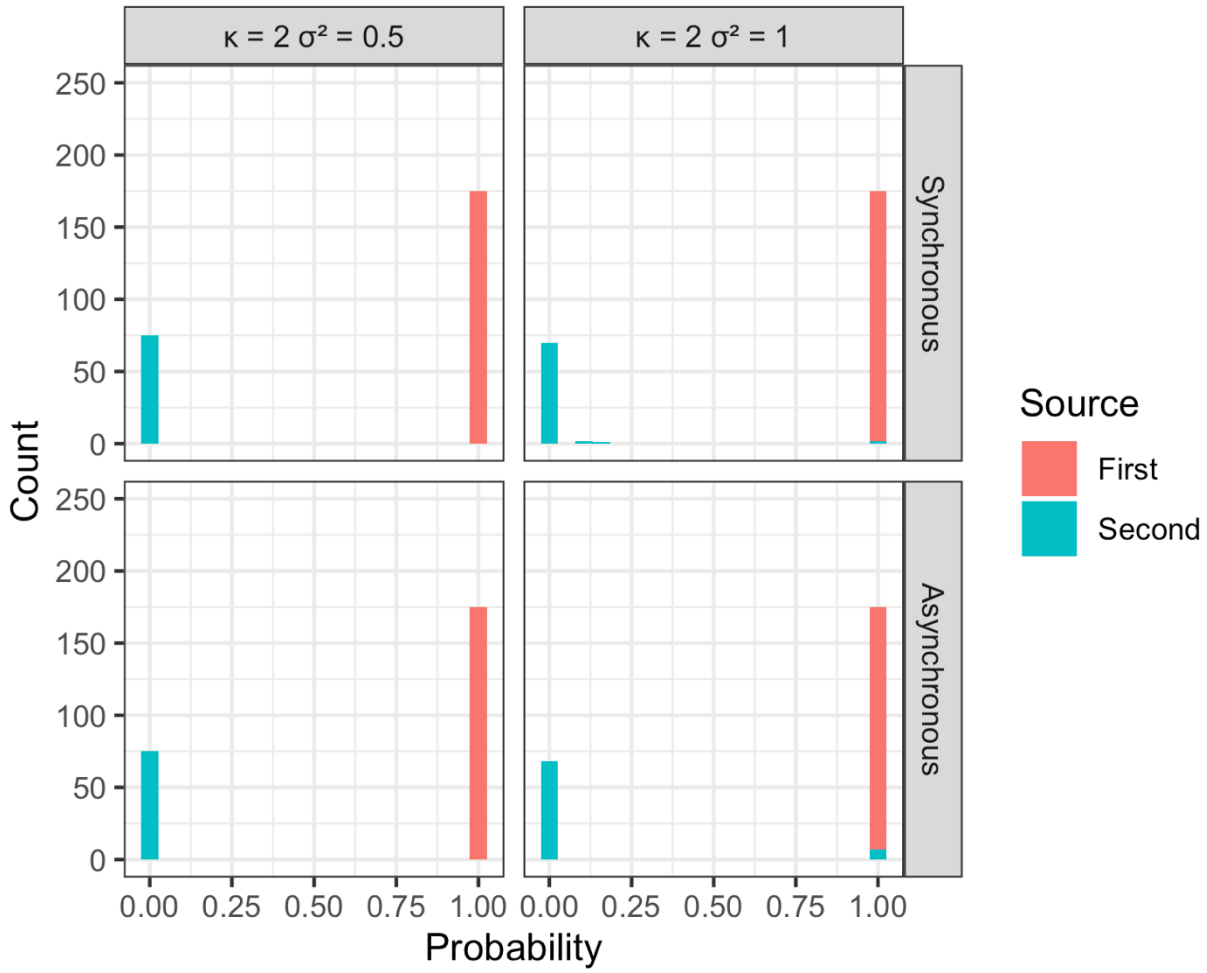


Figure S4.2. The estimated probabilities of the two sources from one simulation for $\kappa = 2$, $\sigma^2 = 0.5$ and 1. The results are shown for synchronous and asynchronous start times.

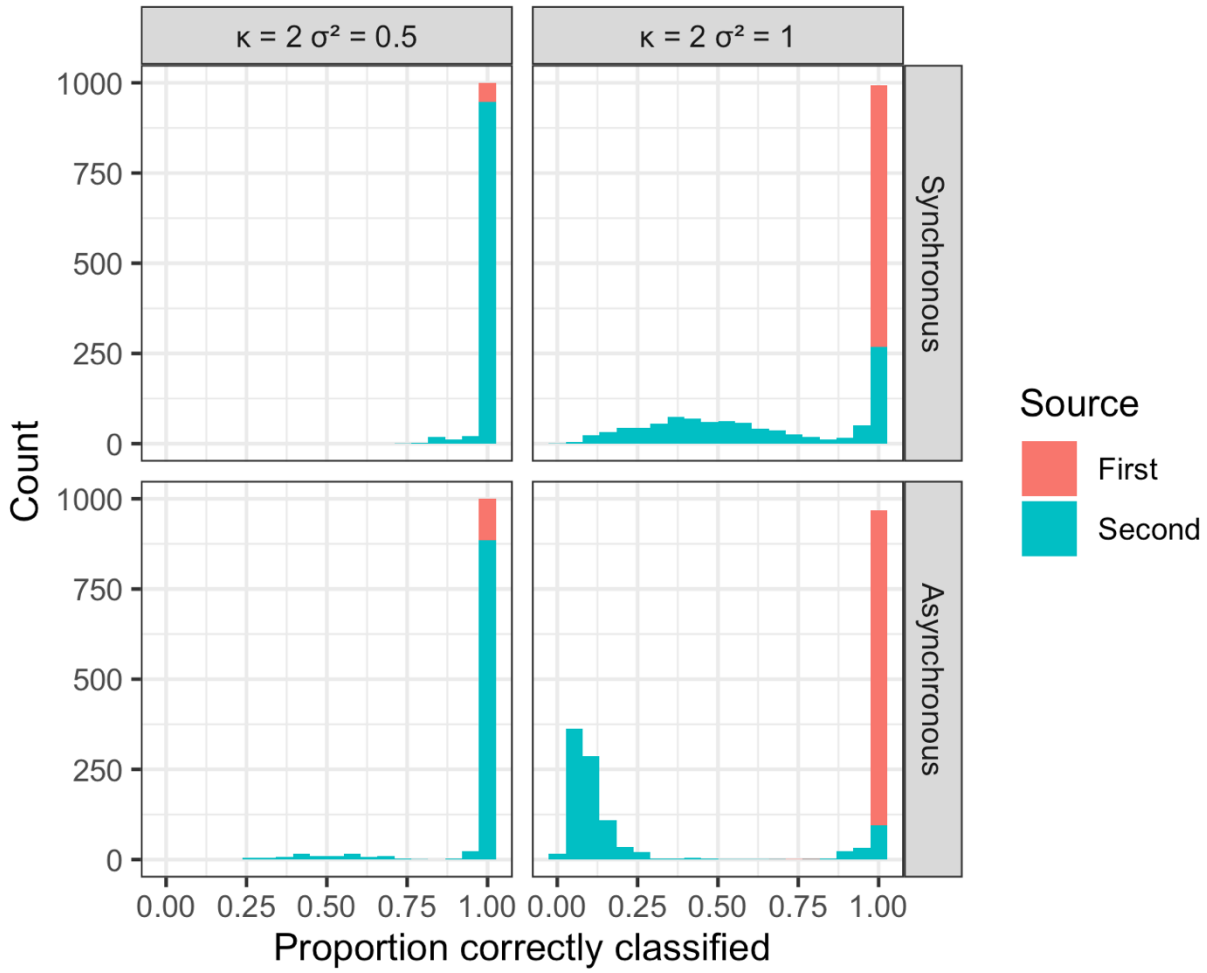


Figure S4.3. The mean proportion of disease correctly assigned to the true source from 1000 simulations for $\kappa = 2$, $\sigma^2 = 0.5$ and 1.

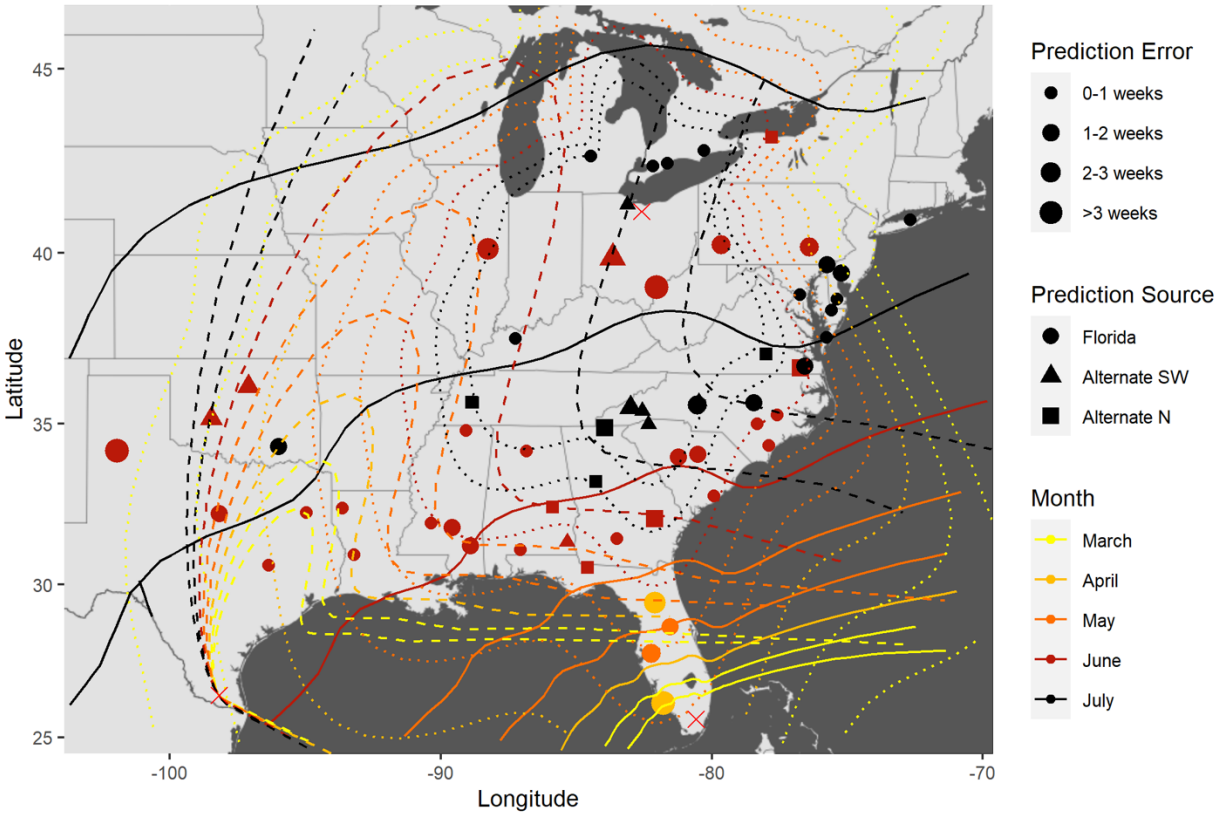


Figure S4.4. Prediction errors and wave front contours from a three-source model fit to the 2009 data. Alternate sources are placed in both the southwest and northern regions; the precise locations are those that yielded the best fits for two-source models.

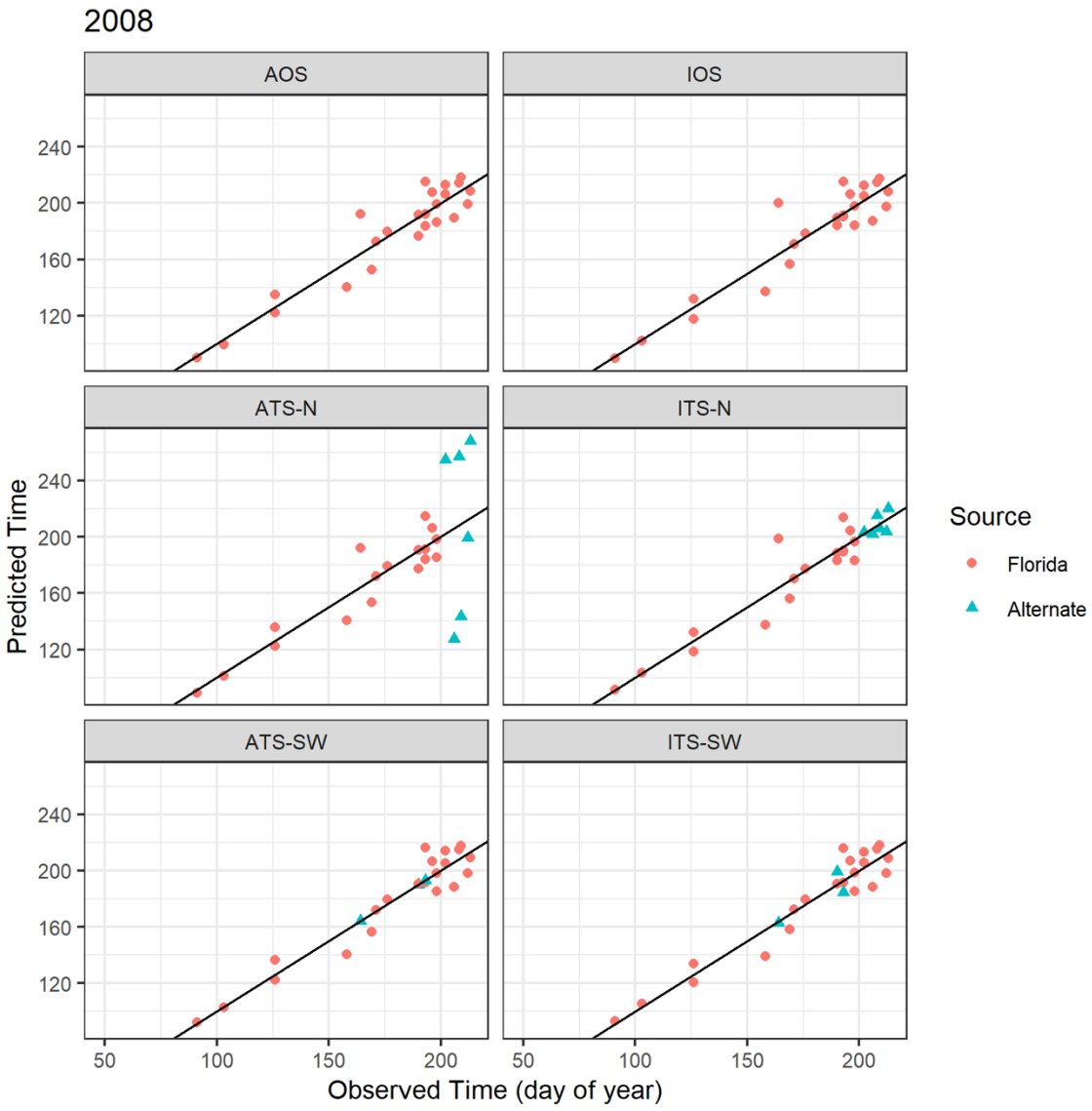


Figure S4.5. Predicted versus observed times of disease occurrence in 2008 for each model shown in Figure 4.6.

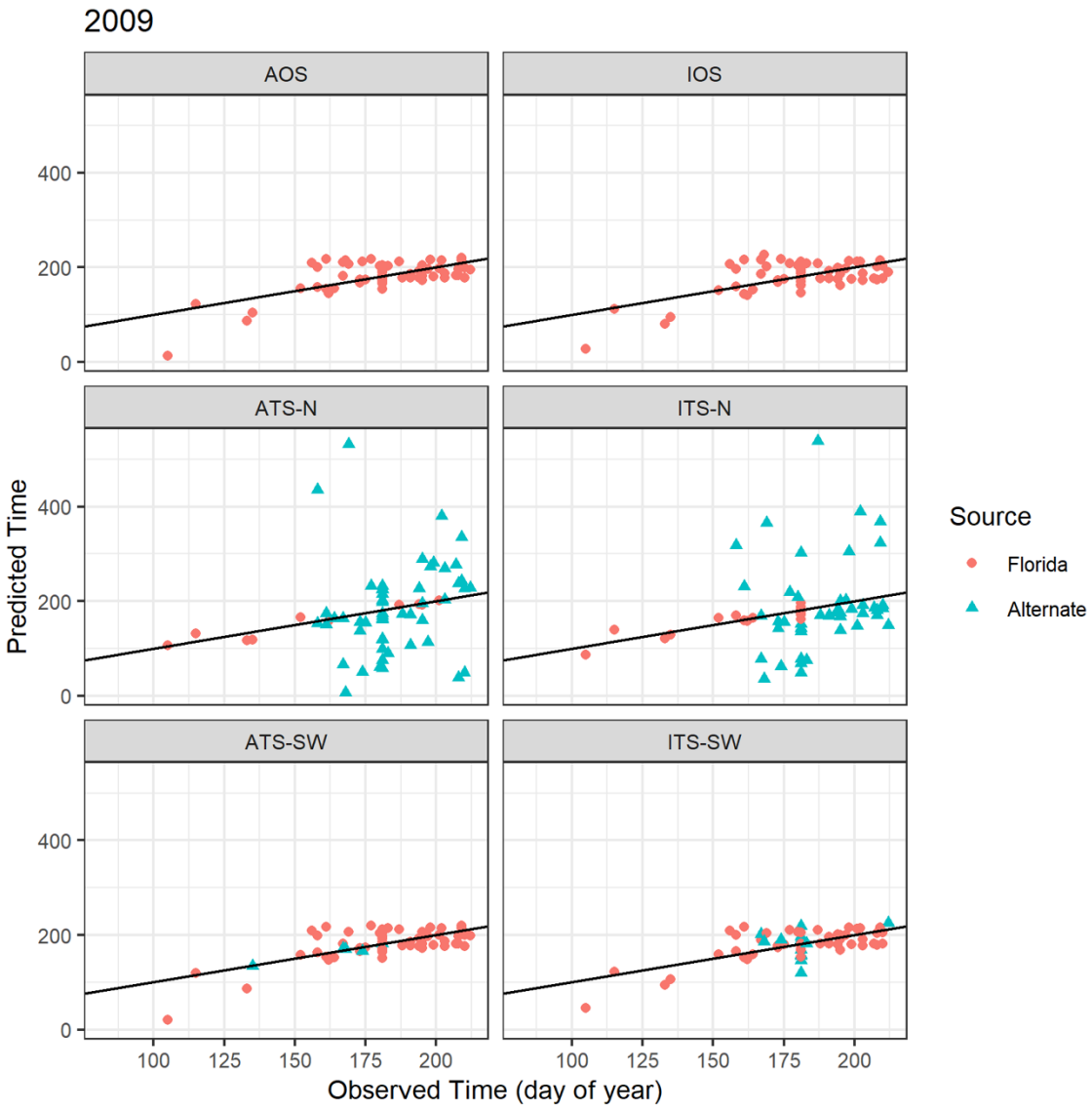


Figure S4.6. Predicted versus observed times of disease occurrence in 2009 for each model shown in Figure 4.7.

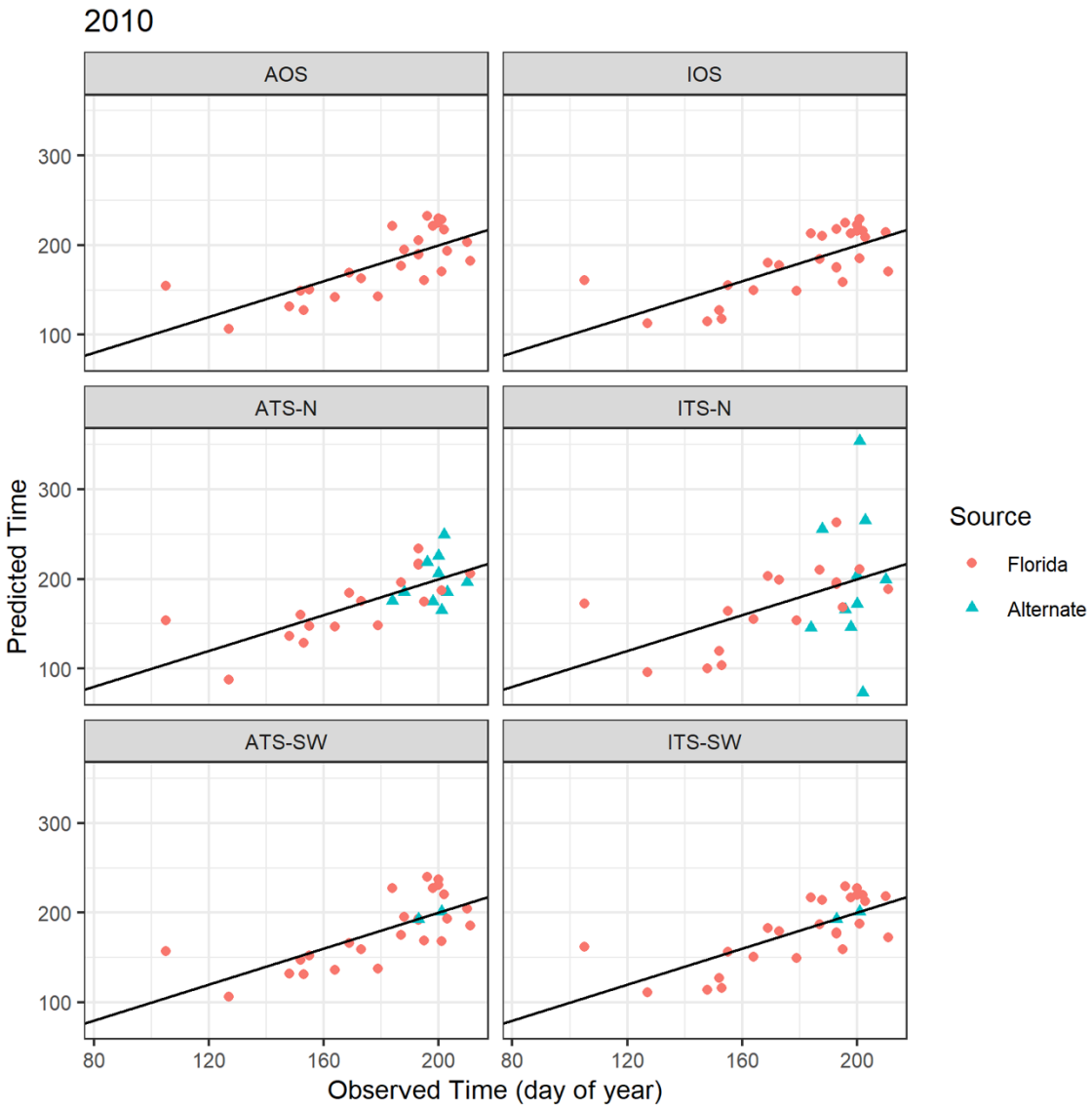


Figure S4.7. Predicted versus observed times of disease occurrence in 2010 for each model shown in Figure 4.8.

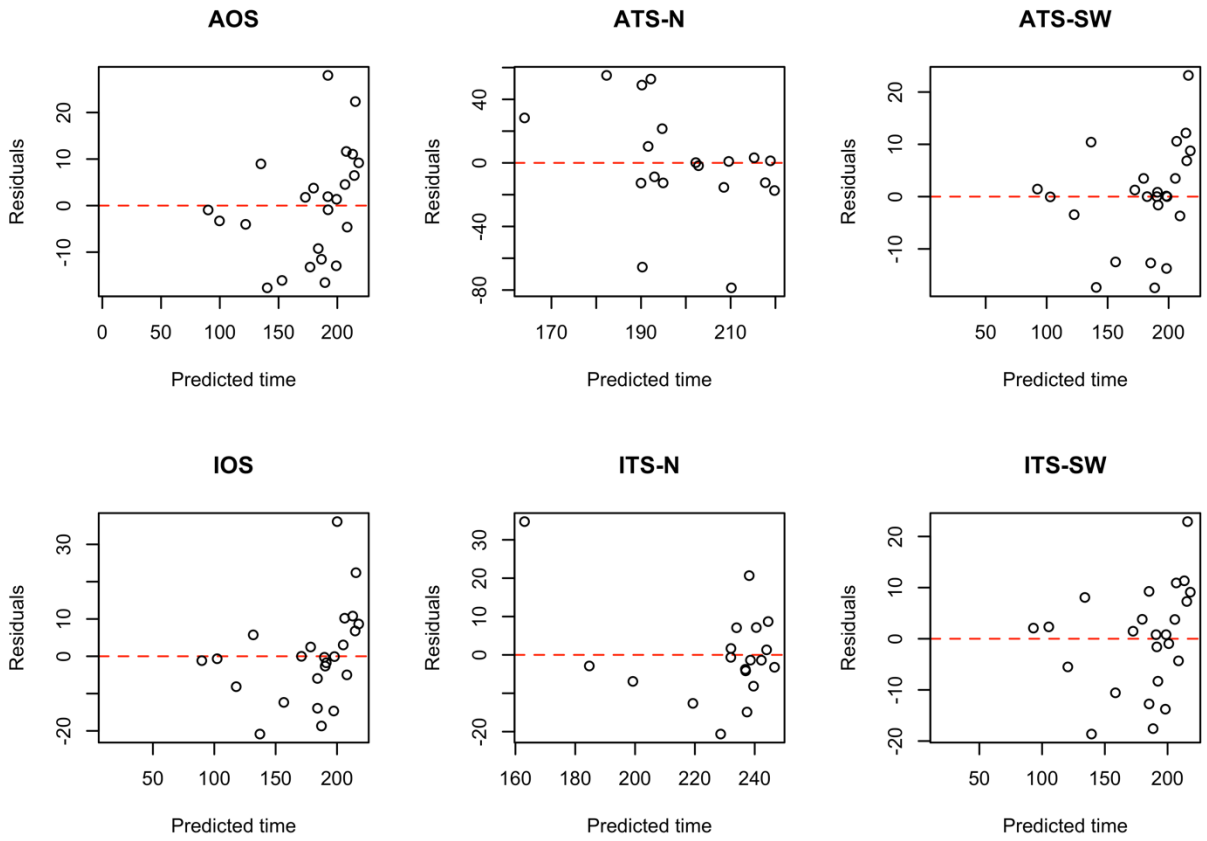


Figure S4.8. Predicted time versus residuals of disease occurrence in 2008 for each model shown in Figure 4.6.

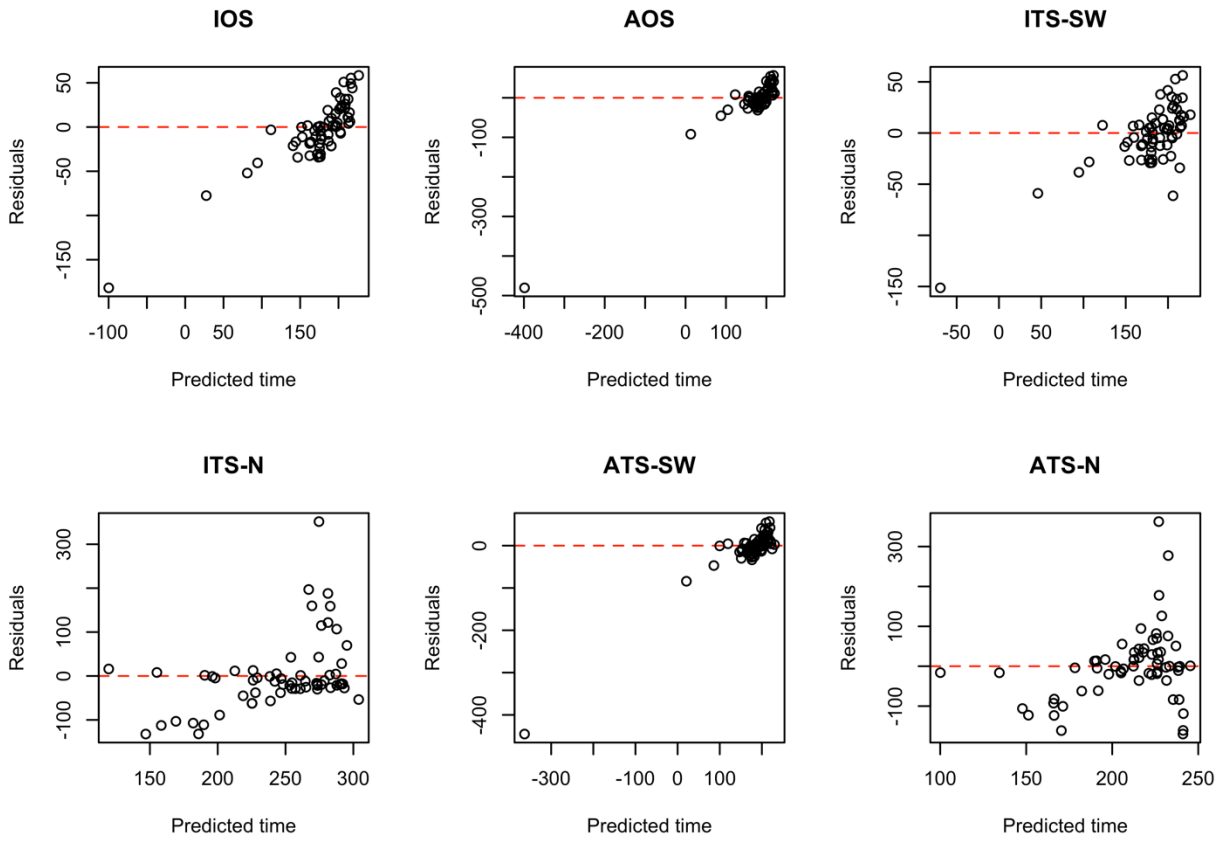


Figure S4.9. Predicted time versus residuals of disease occurrence in 2009 for each model shown in Figure 4.7.

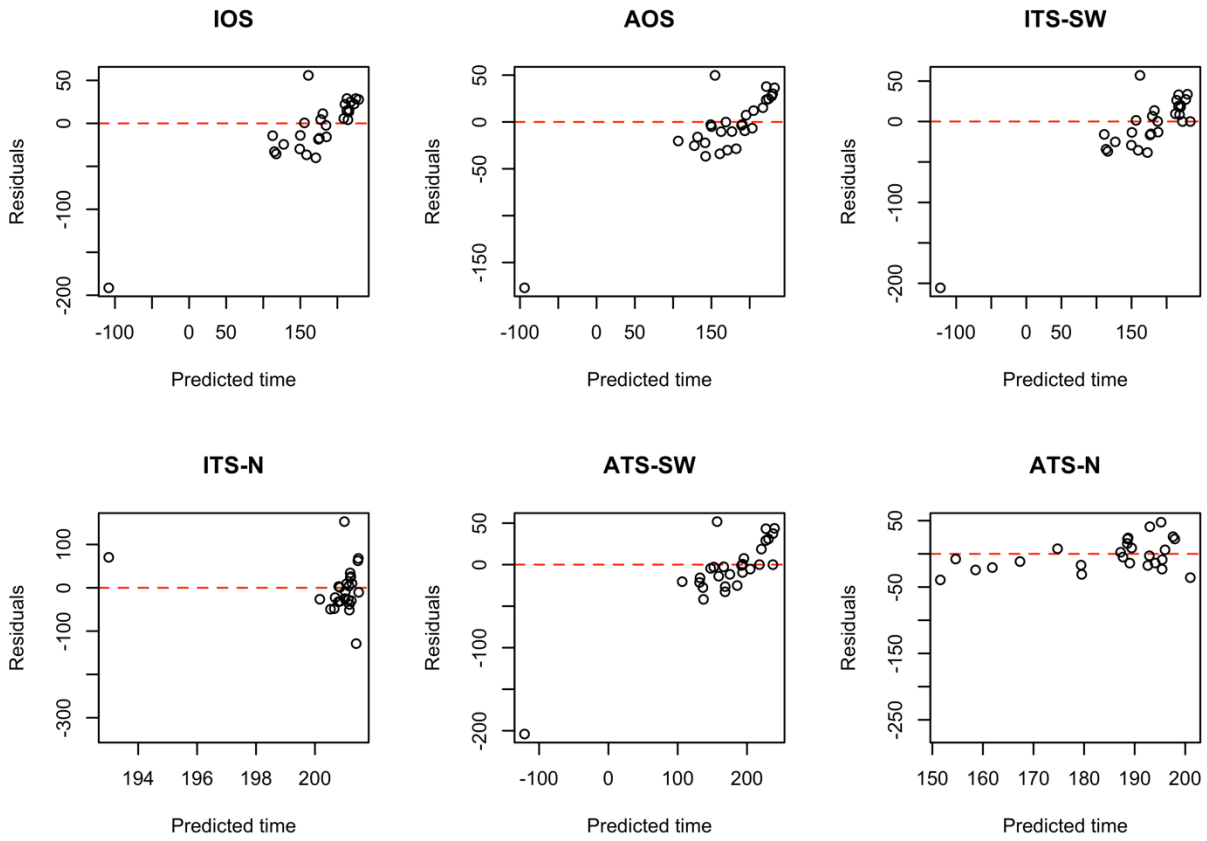


Figure S4.10. Predicted time versus residuals of disease occurrence in 2010 for each model shown in Figure 4.8.

Details on Expectation-Maximization (EM) Algorithm Estimation Procedure

Let $\left(x^{(1)}, \ldots, x^{(K)}\right)$ denote the epidemic source locations for each $k = 1, \ldots,$ with $x^{(k)} \in R^{\mathbb{2}}$. Now an arbitrary location $x \in R^{\mathbb{2}}$ is associated with K sets of polar coordinates $\left(r^{(1)}, \phi^{(1)}\right), \ldots, \left(r^{(K)}, \phi^{(K)}\right)$ where the kth polar coordinate pair indicates the distance $r^{(k)}$and angle $\phi^{(k)}$ to the kth source point $x^{(k)}$. Applying the model framework given in the main paper to each set of coordinates yields the collection of velocity models

$$\log\left(1 + \frac{r^{(k)}}{g_k(\phi^{(k)})}\right) = -M_k t + h_k\left(\phi^{(k)}\right) \qquad k = 1, \ldots, K \tag{1}$$

Now, if multiple sources are present, any given location could be subject to disease exposure from as many as K wavefronts moving simultaneously. Yet, depending on conditions, the movement patterns of the wavefronts, and relative distances to each epicenter, an infection event at any particular time and location is attributable to the different sources with varying probability. In other words, disease at particular locations is more likely due to certain sources rather than others. To accommodate this intuition, a latent process Z is introduced that indicates the relative probabilities of disease associated with each of the K sources, and the collection of models given in Equation (1) describe (r, ϕ, t) conditional on the possible values of Z.

$$\boldsymbol{Z}_{K\times 1} \sim \text{Multinomial}\left(p^{(1)}, \ldots, p^{(K)}\right) \tag{2}$$

$$\log\left(1 + \frac{r^{(k)}}{g_k(\phi^{(k)})}\right) = -M_k t + h_k\left(\phi^{(k)}\right) \quad \text{if } Z_k = 1 \tag{3}$$

We propose an estimation procedure wherein velocity models are fit using regression methods conditional on known g_k. The functions g_k introduce anisotropy in the model. In many applications, known variables drive anisotropy, so it is plausible to estimate g_k from covariate information or secondary data sources.

The velocity models (Equation 1) are fitted conditional on g_k to disease occurrence data (presence or absence) of the form $\mathcal{Y} = \left\{\left(r_i^{(1)}, \phi_i^{(1)}\right), \ldots, \left(r_i^{(K)}, \phi_i^{(K)}\right), t_i\right\}_{i=1}^{n}$ indicating the locations and times of the first observed disease case. For the purpose of exposition, suppose one is fitting only the kth model: consider just the data $\left(r_i^{(k)}, \phi_i^{(k)}, t_i\right)$ and assume $P(Z_i = k) = 1$. Now, adding an offset c_k and Gaussian error term $\epsilon_i^{(k)}$ to Equation 1 yields the statistical model

$$\log\left(1 + \frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) = c_k - M_k t_i + h_k\left(\phi_i^{(k)}\right) + \epsilon_i^{(k)} \tag{4}$$

where $\epsilon_i^{(k)} \sim N(0, \sigma_k^2)$ and $i = 1, \ldots, n$. Under this multiple source situation, the *complete data* can be given as $\mathcal{X} = \left\{\boldsymbol{Z_i}, \left(r_i^{(1)}, \phi_i^{(1)}\right), \ldots, \left(r_i^{(K)}, \phi_i^{(K)}\right), t_i\right\}_{i=1}^{n}$, where $\boldsymbol{Z_i} \in \{0,1\}^K$ for all $i \in \{1, \ldots, n\}$ with only one 1 rest all 0's in each $\boldsymbol{Z_i}$. $\{\boldsymbol{Z_i}\}_{i=1}^{n}$ are considered as the $\{$*unobserved data*$\}$. The source with which the i-th location is principally associated with is given by $\boldsymbol{Z_i}$ by taking the value 1. The unknown parameter set is given by $\Theta = \left\{c_k, M_k, h_k, \sigma_k^2, p^{(k)}\right\}_{k=1}^{K}$. The likelihood of the parameters Θ given the complete data $\mathcal{X}$ is given by

$$\mathcal{L}(\Theta|\mathcal{X}) = \sum_{k=1}^{K} \prod_{i=1}^{n} \left[\varphi\left(\log\left(1 + \frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) - c_k + M_k t_i - h_k\left(\phi_i^{(k)}\right), \sigma_k^2\right)\right]^{\mathbf{1}(Z_{ik}=1)} \tag{5}$$

where $\varphi(x, \sigma^2)$ is the probability density function of a Gaussian random variable with mean zero and variance σ^2 evaluated at value x. The log-likelihood of the parameters Θ given the complete data $\mathcal{X}$ is given by

$$\mathrm{l}(\Theta|\mathcal{X}) = \sum_{k=1}^{K} \mathbf{1}(Z_{ik} = 1) \sum_{i=1}^{n} \left[-\frac{\left(\log\left(1 + \frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) - c_k + M_k t_i - h_k\left(\phi_i^{(k)}\right)\right)^2}{2\sigma_k^2}\right] \tag{6}$$

The expected log-likelihood of the parameters Θ given the complete data $\mathcal{X}$ is given by

(7) $$E[\mathrm{l}(\Theta|\mathcal{X})] = -\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{p_i^{(k)}}{\sigma_k^2}\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) - c_k + M_k t_i - h_k\left(\phi_i^{(k)}\right)\right)^2$$

where, $p_i^{(k)} = E(Z_{ik} = 1|\mathcal{Y})$, the probability of i-th location being principally associated with the k-th source ($i \in \{1, \dots, n\}, k \in \{1, \dots, K\}$). The Expectation-Maximization (EM) algorithm is an iterative algorithm which iterates between the expected log-likelihood and maximizing the expected log-likelihood.

The maximization of the complete log-likelihood in Equation 7 can be broken down into K minimization problems involving weighted least squares problems -

(8) $$\mathrm{l}_k(c_k, M_k, h_k, \sigma_k^2) = -\sum_{i=1}^{n}\frac{p_i^{(k)}}{\sigma_k^2}\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) - c_k + M_k t_i - h_k\left(\phi_i^{(k)}\right)\right)^2$$

The minimization of weighted least squares loss function in Equation (8) leads to estimates of c_k, M_k and h_k given prior estimates of σ_k^2 and $\left\{p_i^{(k)}\right\}_{i=1}^{n}$. Estimates of c_k, M_k and h_k are easily computed using semiparametric regression. Let $s_1(\cdot), \dots, s_B(\cdot)$ denote a set of B basis functions. Now, rewriting Equation 4 we obtain

(9) $$\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) = c_k + (-M_k)t_i + \beta_1^{(k)}s_1\left(\phi_i^{(k)}\right) + \cdots + \beta_B^{(k)}s_B\left(\phi_i^{(k)}\right) + \epsilon_i^{(k)}$$

The weighted least squares (WLS) solution to Equation (9) with weights given by $\left\{p_i^{(k)}/\sigma_k^2\right\}_{i=1}^{n}$, subsequently yields estimates of $\widehat{c_k}$, $\widehat{M_k}$ and $\widehat{h_k} = \sum_b \widehat{\beta_b^{(k)}}\, s_b$ for each $k = 1, \dots, K$. The maximum likelihood estimate of σ_k^2 becomes

(10) $$\widehat{\sigma_k^2} = \sum_{i=1}^{n} \widehat{p_i^{(k)}}\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right) - \widehat{c_k} + \widehat{M_k}t_i - \widehat{h_k}\left(\phi_i^{(k)}\right)\right)^2$$

The estimate of $\left\{p_i^{(k)}\right\}_{i=1}^{n}$ given the estimates $\widehat{c_k}, \widehat{M_k}, \widehat{h_k}, \widehat{\sigma_k^2}$, and $\left\{p_i^{(k)}\right\}_{i=1}^{n}$, the maximum likelihood estimate of $\left\{p_i^{(k)}\right\}_{i=1}^{n}$ becomes

$$\widehat{p_i^{(k)}} = \frac{\varphi\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right)-\widehat{c_k}+\widehat{M_k}t_i-\widehat{h_k}\left(\phi_i^{(k)}\right),\widehat{\sigma_k^2}\right)p_i^{(k)}}{\sum_{k=1}^{K}\varphi\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right)-\widehat{c_k}+\widehat{M_k}t_i-\widehat{h_k}\left(\phi_i^{(k)}\right),\widehat{\sigma_k^2}\right)p_i^{(k)}} \tag{11}$$

Finally, this estimation strategy is extended to the full collection of K models by accounting for the latent variables Z_i that attribute each of the i-th data points to one of the K sources. Formally, the joint likelihood of the data arising from Equations (2) and (3) is maximized with respect to the parameters $p^{(k)} \in R^N$, $\beta_k \in R^{B+2}$, and σ_k^2 for $k = 1, \ldots, K$. The EM algorithm is used to iteratively update estimated multinomial probabilities $\widehat{p_i^{(1)}}, \ldots, \widehat{p_i^{(K)}}$ for each data point in alternation with fitting the regression models in Equation 9 using the estimate $\widehat{p_i^{(k)}}$ as a regression weight for the i-th data point in fitting the k-th model. In detail, the iterations are given by:

1. Initiate $\widehat{p_i^{(k)}}$ as the weight of ith data-point to be associated with kth source, where $\sum_{k=1}^{K}\widehat{p_i^{(k)}} = 1$.
2. Compute/update the estimates $\left(\widehat{c_k}, \widehat{M_k}, \widehat{h_k}, \widehat{\sigma_k^2}\right)_{k=1}^{K}$ by fitting each of the models in Equation 9 with weights $\widehat{p_i^{(k)}}$ for the ith data point and the kth model.
3. Update $\widehat{p_i^{(k)}}$ by

$$\widehat{p_i^{(k)}} = \frac{\varphi\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right)-\widehat{c_k}+\widehat{M_k}t_i-\widehat{h_k}\left(\phi_i^{(k)}\right),\widehat{\sigma_k^2}\right)\widehat{p_i^{(k)}}}{\sum_{k=1}^{K}\varphi\left(\log\left(1+\frac{r_i^{(k)}}{g_k\left(\phi_i^{(k)}\right)}\right)-\widehat{c_k}+\widehat{M_k}t_i-\widehat{h_k}\left(\phi_i^{(k)}\right),\widehat{\sigma_k^2}\right)\widehat{p_i^{(k)}}} \tag{12}$$

where $\varphi(x, \sigma^2)$ is the probability density function of a Gaussian random variable with mean zero and variance σ^2 evaluated at value x.

4. Repeat steps 2-3 until convergence.

A simple heuristic for the initialization step is to use as $\widehat{p_i^{(k)}}$ the estimated probabilities obtained by logistic regression of an indicator of whether the kth source is closest on the variables $r^{(1)}/\widehat{g_1}\left(\phi^{(1)}\right), \ldots, r^{(K)}/g_k\left(\phi^{(K)}\right)$. We note that an isotropic model with one or many sources can be recovered within this framework as a special case by fixing $g_k(x) = 1/2$ for $x \in [0,2\pi]$, with the consequence that $h_k \equiv 0$.

CHAPTER 5

Conclusion

A better understanding of the impacts of epidemic control on cucurbits is vital for developing efficient control programs. Without this knowledge, control programs risk being ineffective at informing timely initial fungicide application and reducing epidemic invasion. The current platform for monitoring, predicting, and communicating the risk of CDM outbreaks in the eastern United States has been beneficial; however, it is expensive to maintain, and the resources are often limited. In this dissertation, studies are conducted to identify ways of reducing the spread of cucurbit downy mildew to provide improved guidance for the current decision support platform. This work is highly interdisciplinary, borrowing techniques from mathematics, statistics, plant pathology, and network science.

Chapter 2 of this dissertation discusses the use of static and dynamic networks to characterize CDM dynamics. Based on the assumption that the field connectivity influences CDM spread in time and space, networks for dispersal of *Pseudoperonospora cubensis* and the spread of CDM are characterized and found to sensitive to the choice of parameters and thresholds for construction. Most significantly, it is shown that dynamic networks can facilitate visualization of the prominent pathways of disease spread, and areas that are likely to act as sources and promote the spread of CDM are identified. When complemented with disease scouting efforts, these results could be used as a decision support system to inform uncertain situations with regards to locations of initial disease outbreaks in the eastern United States.

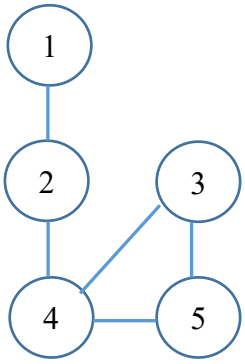
In Chapter 3, the effects of combining centrality measures, frequency of infection, and probability of field infection to identify the locations for disease surveillance and management are quantified. More importantly, it is shown that removing locations identified as the most important based on betweenness centrality greatly limited the risk of disease spread. These locations located in Maryland, North Carolina, Ohio, South Carolina, and Virginia may inform surveillance and

strategies for controlling CDM in the eastern United States. In addition, these locations can be targeted for fungicide treatment to slow down the spread of CDM to disease-free neighboring cucurbit fields.

Chapter 4 explores two existing phenomenological models and extends them by incorporating anisotropy and multiple inoculum sources. Based on the data analysis from 2008-2010, there is a small but consistent reduction in errors associated with incorporating anisotropy into the model regardless of the number of sources, a reduction in errors in certain years associated with incorporating an alternate inoculum source in the southwest. However, there is no reduction in errors associated with incorporating inoculum sources in the north. These results strongly suggest that the initial inoculum for CDM outbreaks in the continental United States is primarily from overwintering sources in the southern United States that are typical sub-tropical in nature.

APPENDIX

Eigenvector Centrality



Iteration 1

$$\begin{bmatrix}0&1&0&0&0\\1&0&0&1&0\\0&0&0&1&1\\0&1&1&0&1\\0&0&1&1&0\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}=\overset{A}{\begin{bmatrix}1\\2\\2\\3\\2\end{bmatrix}}\equiv\overset{B}{\begin{bmatrix}0.213\\0.426\\0.426\\0.639\\0.426\end{bmatrix}}$$

Normalized value

$\sqrt{1^2+2^2+2^2+3^2+2^2}=4.69$ B = A/4.69

Iteration 2

$$\begin{bmatrix}0&1&0&0&0\\1&0&0&1&0\\0&0&0&1&1\\0&1&1&0&1\\0&0&1&1&0\end{bmatrix}\begin{bmatrix}0.213\\0.426\\0.426\\0.639\\0.426\end{bmatrix}=\overset{A}{\begin{bmatrix}0.426\\0.852\\1.065\\1.278\\1.065\end{bmatrix}}\equiv\overset{B}{\begin{bmatrix}0.195\\0.389\\0.486\\0.584\\0.486\end{bmatrix}}$$

Normalized value

$\sqrt{0.426^2+0.852^2+1.065^2+1.278^2+1.065^2}=2.19$

B = A/2.19

Iteration 3

$$\begin{bmatrix}0&1&0&0&0\\1&0&0&1&0\\0&0&0&1&1\\0&1&1&0&1\\0&0&1&1&0\end{bmatrix}\begin{bmatrix}0.195\\0.389\\0.486\\0.584\\0.486\end{bmatrix}=\overset{A}{\begin{bmatrix}0.389\\0.779\\1.070\\1.361\\1.070\end{bmatrix}}\equiv\overset{B}{\begin{bmatrix}0.176\\0.352\\0.484\\0.616\\0.484\end{bmatrix}}$$

Normalized value

$\sqrt{0.389^2+0.779^2+1.07^2+1.361^2+1.07^2}=2.21$

B = A/2.21

Iteration 4

$$\begin{bmatrix}0&1&0&0&0\\1&0&0&1&0\\0&0&0&1&1\\0&1&1&0&1\\0&0&1&1&0\end{bmatrix}\begin{bmatrix}0.176\\0.352\\0.484\\0.616\\0.484\end{bmatrix}=\overset{A}{\begin{bmatrix}0.352\\0.792\\1.100\\1.320\\1.100\end{bmatrix}}\equiv\overset{B}{\begin{bmatrix}0.159\\0.358\\0.497\\0.278\\0.498\end{bmatrix}}$$

Normalized value

$\sqrt{0.352^2+0.792^2+1.1^2+1.32^2+1.1^2}=2.21$ B = A/2.21 converges

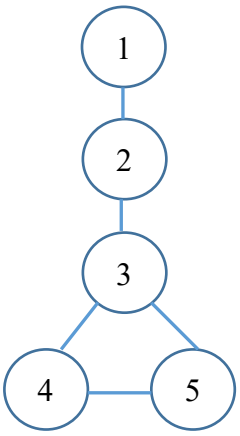
Eigenvector centrality

$$\overset{C}{\begin{bmatrix}0.176\\0.352\\0.484\\0.616\\0.484\end{bmatrix}}\overset{D}{\begin{matrix}1\\2\\3\\4\\5\end{matrix}}$$

C - Eigenvector centrality
D - Node

This example illustrates the calculation of eigenvector centralities for five nodes in a network (courtesy of Dr. Natarajan Meghanathan's lecture notes on centrality measures).

Betweenness Centrality



For each node *k*,

1. Determine the levels i.e., Level 0 = node k, Level 1 = nodes that are 1 hop away, Level 2 = nodes that are 2 hops away etc.
2. Count the shortest paths from a node *i* to a node *j* through node k (numerator). Count other shortest paths (denominator).
3. BWC is the sum of the number of the fractions

BWC for node 2	BWC for node 3
(1,3) = 1/1	(1,4) = 1/1
(1,4) = 1/1	(1,5) = 1/1
(1,5) = 1/1	(2,5) = 1/1
	(2,4) = 1/1
3	4

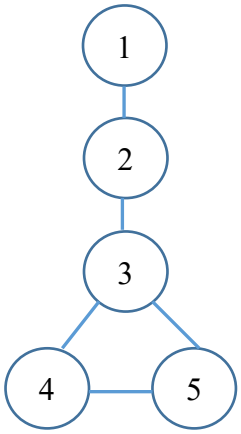
Node	BWC	Rank
1	0	-
2	3	2
3	4	1
4	0	-
5	0	-

(1,3) represents the shortest path between 1 and 3 through node 2 (level 1) = 1. This is the only path connecting 1 and 3 (total path = 1). Therefore (1,3) = 1/1. This simple example illustrates the calculation of betweenness centralities for five nodes in a network. Node 3 has the highest betweenness score and is thus the most central in the network. Removing node 3 will break the network.

Closeness Centrality

For each node *k*,

1. Count the number of hops from k to other nodes
2. Get sum of (1) (Row Sum)
3. Find the reciprocal of (2) (Closeness)



Node	1	2	3	4	5	Row Sum	Closeness
1	0	1	2	3	3	9	0.1111
2	1	0	1	2	2	6	0.1667
3	2	1	0	1	1	5	0.2000
4	3	2	1	0	1	7	0.1429
5	3	2	1	1	0	7	0.1429

Rank	Node	Score
1	3	0.2000
2	2	0.1667
3	4	0.1429
4	5	0.1429
5	1	0.1111

This is a simple example to illustrate the calculation of closeness centralities of five nodes in a network. Node 3 has the highest closeness score and is thus the most central in the network.