Accounting for model uncertainties in statistical forecasts of wildfire parameters

Harry Richard Podschwit

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Ernesto Alvarado-Celestin, Co-Chair

Alison Cullen, Co-Chair

Narasimhan Larkin

Elizabeth Steel

Program Authorized to Offer Degree:

Quantitative Ecology and Resource Management

University of Washington

Abstract

Accounting for model uncertainties in statistical forecasts of wildfire parameters

Harry Richard Podschwit

Chairs of the Supervisory Committee:
Ernesto Alvarado-Celestin
School of Environmental and Forest Sciences

Alison Cullen
Evans School of Public Policy and Governance

Gauging the magnitude of model uncertainty and incorporating model uncertainty into predictions is of critical importance when models are used to inform wildfire-related decisions, where ignoring potential risks threaten human health, property, and the environment. Although techniques exist for addressing model uncertainty, these uncertainties are commonly ignored in most analyses. In this dissertation, I will evaluate the effects of model uncertainty on statistical predictions of wildfire activity in multiple contexts and propose techniques to incorporate these uncertainties into predictions. I will determine how uncertainty in the choice of predictive model and climate model influence forecasts of very-large fire activity in the second half of the 21st century, and integrate this uncertainty using a novel Bayesian model averaging approach to

produce robust predictions. I find that when these model uncertainties are accounted for, that one may conclude, across the suite of model choices, that the frequency of very-large wildfires should be expected to increase in most regions of the United States if climate changes are not mitigated. The effects of model uncertainty will also be explored in the context of predicting final wildfire size for individual fires that have no yet finished growing. Specifically, I will gauge how the choice of utility function and the inclusion of growth information that is unavailable early in the wildfire's life alters the predictive ability of statistical models of final fire size and the stability of the model structure. I find that predictions of fire size can drastically change when new utility functions are considered, particularly in models that use growth information. I also find that the covariates used in the best model are sensitive to the choice of utility function, and that no single model is likely to optimally address the preferences of all wildfire-related decisionmakers they are intended to inform. The results of this analysis that (1) the preferred model will often change when new performance measures are considered, and (2) that the preferred model may change over time. I also present a method of integrating the model uncertainties associated with time-varying covariates and ill-defined utility functions into a single predictive distribution using Bayesian model averaging. I find that this novel model averaging approach generally improves predictive performance across a number of performance measures compared to the individual models contained within it. I discuss how the novel methods developed can be applicable to other forecasting applications and how they might allow wildfire professionals make better decisions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank the AirFire team for their assistance from the beginning of my research. I would also like to specifically thank my colleagues Brian Potter, Linda Mearns, Travis Axe, Nat Goodby, Zach Kearl, Seth McGinnis, Lee Kessenich, Melissa Bukovsky, Fabiola Chavez, Colton Miller, Paulina llamas Casillas, and Lesly Franco. This research was funded in part by AirFire and the National Center for Atmospheric Research.

DEDICATION

This dissertation is dedicated to all of my family: Nancy Podschwit, Megan Marks, Harlan Podschwit, Katie Podschwit, and Alex Podschwit.

Chapter 1. INTRODUCTION

Overconfidence is a near-ubiquitous tendency in humans (Morgan and Henrion 1992). For example, most people when asked to provide an interval that contains some uncertain quantity (e.g. the population of a random city) with a 90 percent probability, will provide intervals that contain the truth much less frequently than 90 percent (Teigen, K. H., & JØrgensen, M 2005). Although the quantity in this case is definite and to at least some individuals completely known, many quantities that are of interest are completely unknown. For instance, an individual who is getting married or starting a business does not know in the beginning the process whether the project will be successful. In this case too, people's predictions tend to be overly-optimistic, as a wealth of studies have shown to be the case. Specifically, people's predictions of future outcomes are likely to be biased in such a way that the desirable outcome (e.g. my marriage/business will be successful) appear more likely than statistics indicate (Korobkin and Guthrie 2003). In many cases, the material risks associated with such overconfidence are low. If no wagers are made, an incorrect belief that an individual coin flip is more likely than chance to result in heads because the previous several flips have been tails is harmless. However, when the costs associated with the outcomes increase, it is important to minimize these overconfident tendencies.

With these abstractions in mind, I will note now that in the specific case of wildfire, predictions are used to inform decisions that have very costly outcomes. These costs can come from a variety of dimensions. Fighting wildfire costs money, and one major source of direct economic costs arise from decisions to suppress fires, particularly large fires (González-Cabán 1983,

Stephens et al., 2014). These severe economic costs can also arise from property losses (Barrett 2018), which in some cases may be much higher if firefighting is restrained. In addition to these direct economic costs, there are a suite of indirect economic costs of wildfire as well. These indirect costs can include damages from post-fire hazards, rehabilitation expenses, lost tax, and lost business revenue from community evacuations (Dale 2009, Neary et al., 2003, Peppin et al., 2011, Beverly and Bothwell 2011, Beverly et al., 2011). Health costs can arise as well, and the flames of a wildfire are an obvious source of human injury and death. However, the same outcomes can be realized via less direct processes. For instance, wildfires produce smoke that can linger for months (Stephens et al., 2014) and affect humans far from active burning sites (Forster et al., 2001, Val Martin et al., 2013). Wildfire smoke inhalation can negatively impact public health (Reid et al., 2016, Moeltner et al., 2013), and safety (Achtemeier 2009, Stephens et al., 2014). Environmental costs of wildfire are also noteworthy. Wildfires release large amounts of greenhouse gases (Liu et al. 2014), facilitate the establishment of invasive species (Crawford et al., 2001), promote the loss of ecosystem services (Rocca et al., 2014), and can cause long-term alterations to forest structure (Haffey et al., 2018), and accelerate global warming via black carbon deposition (Larkin et al. 2014).

Costs of wildfire - in a broad sense such as suppression costs, property damages, ecosystem services, etc. - can sometimes be lessened by decision-makers. For example, evacuating a certain population during an individual fire might mitigate the health risks of wildfire (Moritz et al. 2014), as could distributing important information to affected populations, or canceling nearby events (Sugerman et al. 2012). Firefighting can protect property and other values-at-risk from wildfire, and strategic use of fire can lessen future risks of wildfire (North et al. 2015). Electrical

companies could turn off power to prevent ignitions from power lines. Fire growth could be slowed or stopped in critical area by prepositioning firefighters or constructing fire breaks. Law enforcement could enforce mandatory evacuation orders well in advance of hazardous conditions.

The variability in wildfire costs can be high. A wildfire burning in rangelands will not have the same costs as one that burns in a densely populated urban area. At least some of this variability in costs can be accounted for using certain wildfire parameters. For instance, average firefighting costs increase with fire size (González-Cabán 1983) and firefighting effort has been observed to correlate with the number of fires (Podschwit and Cullen 2020). Ecological recovery effort will correlate with the amount of area burned at high-severity (Pelletier and Orem 2014, Keane et al. 2009), and duration can further refine estimates of firefighting costs (Gebert and Black 2012). An abridged list of the types of wildfire parameters that are relevant to certain decision-makers is shown in (Table 1.1)

Table 1.1. Wildfire characteristics and the kinds decisions they may inform.

Wildfire characteristic	Relevant decision-making contexts
Burn area	Firefighting costs (González-Cabán 1983) Evacuations (Beverly and Bothwell 2011)
Growth rate	Firefighter safety (Viegas and Simeoni 2011)
Duration	Firefighting costs (Gebert and Black 2012) Initialization of wildfire growth models (Finney 1998)
Perimeter	Fireline production objectives (Katuwal et al. 2016)
Fire severity	Burn area emergency response (Robichaud et al. 2014)
Simultaneous wildfire	Resource demand (Bednar et al. 1990, Podschwit and Cullen 2020)

In some cases, the mitigation costs (e.g. firefighting) could be higher than the dollar damages of wildfire, implying that such interventions are not always cost effective. If future wildfire parameters that correlated with these damages were known with certainty, then decisionmakers could be more confident that interventions intended to mitigate these damages were cost effective. For instance, if firefighters knew in advance that a wildfire would not grow into a densely populated urban community, then immediate firefighting costs could be lessened by adopting a less aggressive suppression strategy. In this case also, long-term firefighting costs may be lessened by the removal of hazardous fuels. Note that costs may be conflict. For instance, firefighters may desire to extinguish a wildfire that is threatening a densely populated urban community, but this suppression of wildfire may be at the expense of realizing some ecological benefits of wildfire. However, the future costs can never be known with certainty and must be assessed probabilistically. This is accomplished through the use models which allow decisionmakers to make informed guesses about the future. However, the use of models themselves introduce an opportunity for overconfidence.

Overconfidence arises when uncertainties are ignored and one uncertainty that is commonly ignored is unknowns regarding the structure of model used to generate predictions. There are often multiple plausible model structures to consult and it is not clear which, or if only one, model should be used. For instance, predictive models of wildfire occurrence and spread can be stochastic or deterministic (Taylor et al. 2013) and selecting one ignores the possibility that the other modeling category is potentially informative. Even within these modeling categories, there is potential to ignore potentially relevant models. For instance, there are many factors that may be relevant to wildfire prediction. Wildfire activity is mediated through a number of distinct

processes that include weather (Flannigan et al. 2009), atmospheric processes (Brotak 1977), vegetation (Meyn et al. 2007, Bradley et al. 2016), and anthropogenic factors (Syphard et al. 2017, Syphard et al. 2007, Brotons et al. 2013). Selecting which factors to include in a predictive model is not always obvious or a matter of choice. Dryness is often identified as an important factor in mediating wildfire activity (Barbero et al. 2014), but multiple proxies could still be used to represent this factor (Zargar et al. 2011). Additionally, even when the factors relevant to prediction are known or agreed upon, the model structure itself might be indeterminate. There are multiple model structures that could be used to represent the relationships between wildfire and the environment, including generalized linear models (Littell et al. 2009, Barbero et al. 2014, Stavros et al. 2014, Guo et al. 2015), and machine learning approaches (Rodrigues and De la Riva 2014, Zou et al. 2019). Even if the universe of all possible models were somehow tractable, selecting the most appropriate one requires defining a utility function to gauge model quality, which is also indeterminate. The utility function to gauge model quality may differ across decision makers, as some may care about one aspect of the model (minimizing errors) and another decision maker may care about another (maximizing recall). Because no one model is likely to be best at all things, the choice of utility function can change which model is selected. Rankings of model quality commonly vary across performance measures (Willmott 1982, Morgan and Henrion 1992) and rankings of wildfire models are likely to similarly vary across the variety of performance measures proposed to assess the model quality (Cruz and Alexander 2013, Filippi et al. 2014). Even when a performance measure is agreed upon, competing models may have near identical predictive performance (Johnson and Omland 2004), and it can be unclear if any designation of “best” model will stay the same when new data are considered.

Despite the many sources of model uncertainty, by far the most common approach to fire modeling is to ignore uncertainty and focus on only one model (Catry et al. 2007, Stavros et al. 2014, Hoeting et al. 1999). This single model approach is risky because results from the unconsidered models may drastically differ from the selected one. Hence, uncertainty in wildfire forecasts is artificially low, which can have catastrophic consequences if the results inform important decisions (Draper 1995). Incorporating structural uncertainties into model predictions can safeguard decision makers against unjustified and unnecessary risk tolerance. One way to deal with multiple plausible models is to compare results from individual model candidates to assess the sensitivity of predictions to alternative candidates or generate scenarios (Morgan and Henrion 1992, Littell et al. 2011). However, when the number of candidate models is large, presenting results of individual models can be onerous. Metamodeling, creating model of models, then becomes an appealing alternative to integrating structural uncertainty into predictions.

Bayesian model averaging (BMA) is a flexible and commonly used method of metamodeling. Within a BMA framework, model weights, which are assumed to be uncertain, are used to combine covariates or predictions from multiple sources into a single probability distribution. Model weights are represented using the posterior distribution, which is a probability distribution representing one's belief about the model parameters conditional on the observed data. The resulting distribution of model weights can be used to generate histograms of predictions of the quantities relevant to decision making, while also accounting for structural uncertainties that would be ignored if a single model were used. BMA simultaneously has been shown to lessen

many of the risks of traditional model selection techniques and to improve performance across a variety of metrics (Raftery and Zheng 2003).

In this dissertation I will gauge the prevalence of model uncertainties in predicting wildfire occurrence and size and identify multiple sources of model uncertainty. In Chapter 2, I will incorporate multiple sources of model uncertainty into projections of very-large fire activity in the second half of the 21st century. In Chapter 3, I will explore how changes to the utility function, as well as changes in information availability over time, can lead to changes in the optimal structure of statistical models of wildfire size. I will also describe the sensitivity of fire size predictions to these changes model structure. I will close this dissertation by describing a metamodel that can incorporate model uncertainties, that arise from the sources described in Chapter 3, into a single predictive model of fire size.

1.1 REFERENCES

Achtemeier, G.L. On the formation and persistence of superfog in woodland smoke. *Meteorological Applications*. 2009, 16, 215–225.

Barbero, R., Abatzoglou, J.T., Steel, E.A., & Larkin, N.K. (2014). Modeling very large-fire occurrences over the continental United States from weather and climate forcing. *Environmental Research Letters*. 9(12), 124009.

Barrett, K. The Full Community Costs of Wildfire. Headwaters Economics. Available online: https://headwaterseconomics.org/wp-content/uploads/full-wildfire-costs-report.pdf (accessed on 14 December 2018).

Bednar, L.F., Mees, R., & Strauss, D. (1990). Fire suppression effectiveness for simultaneous fires: An examination of fire histories. *Western Journal of Applied Forestry*. *5*(1), 16-19.

Beverly, J.L., & Bothwell, P. Wildfire evacuations in Canada 1980–2007. (2011) *Natural Hazards*. 59(1), 571–596.

Beverly, J.L., Flannigan, M.D., Stocks, B.J., & Bothwell, P. The association between Northern Hemisphere climate patterns and interannual variability in Canadian wildfire activity. *Canadian Journal of Forest Research*. 2011, 41, 2193–2201.

Bradley, B.A., Curtis, C.A., & Chambers, J.C. (2016). Bromus response to climate and projected changes with climate change. In Exotic Brome-Grasses in Arid and Semiarid Ecosystems of the Western US (pp. 257-274). Springer International Publishing.

Brotak, E.A., & Reifsnyder, W.E. (1977). An investigation of the synoptic situations associated with major wildland fires. *Journal of Applied Meteorology*. *16*(9), 867-870.

Brotons, L., Aquilué, N., De Cáceres, M., Fortin, M.J., & Fall, A. (2013). How fire history, fire suppression practices and climate change affect wildfire regimes in Mediterranean landscapes. *PLOS one*. 8(5), e62392.

Calkin D.E., Gebert K.M., Jones J.G., Neilson R.P. (2005) Forest service large fire area burned and suppression expenditure trends, 1970–2002. *Journal of Forestry.* **103**, 179–183. doi:10.1093/JOF/103.4.179

Catry, F.X., Rego, F.C., Bação, F.L., & Moreira, F. (2010). Modeling and mapping wildfire ignition risk in Portugal. *International Journal of Wildland Fire*. 18(8), 921-931.

Crawford, J.A., Wahren, C.H., Kyle, S., & Moir, W.H. Responses of exotic plant species to fires in Pinus ponderosa forests in northern Arizona. *Journal of Vegetation Science*. 2001, 12, 261–268.

Cruz, M.G., & Alexander, M.E. (2013). Uncertainty associated with model predictions of surface and crown fire rates of spread. *Environmental Modelling & Software*. 47, 16-28.

Dale, L. (2009) The True Cost of Wildfire in The Western US; Western Forestry Leadership Coalition: Denver, CO, USA.

Filippi, J.B., Mallet, V., & Nader, B. (2014). Representation and evaluation of wildfire propagation simulations. *International Journal of Wildland Fire*. 23(1), 46-57.

Finney, M.A. (1998). FARSITE: Fire Area Simulator-model development and evaluation. *Res. Pap. RMRS-RP-4, Revised 2004. Ogden, UT: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 47 p.*, *4*.

Flannigan, M.D., Krawchuk, M.A., de Groot, W.J., Wotton, B.M., & Gowman, L.M. (2009). Implications of changing climate for global wildland fire. *International Journal of Wildland Fire*. 18(5), 483-507.

Forster, C., Wandinger, U., Wotawa, G., James, P., Mattis, I., Althausen, D., Simmonds, P., O'Doherty, S., Jennings, S.G., Kleefeld, C., & Schneider, J. (2001) Transport of boreal forest fire emissions from Canada to Europe. *Journal of Geophysical Research: Atmospheres*. 106(D19), 22887–22906.

Gebert, K.M., & Black, A.E. (2012). Effect of suppression strategies on federal wildland fire expenditures. *Journal of Forestry*. 110(2), 65-73.

González-Cabán, A. Economic Cost of Initial Attack and Large-Fire Suppression. (1983). USDA Forest Service General Technical Report PSW-068; U.S. Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station: Berkeley, CA, USA, 7p.

Guo, F., Wang, G., Innes, J. L., Ma, X., Sun, L., & Hu, H. (2015). Gamma generalized linear model to investigate the effects of climate variables on the area burned by forest fire in northeast China. *Journal of Forestry Research*. 26(3), 545-555.

Haffey, C., Sisk, T.D., Allen, C.D., Thode, A.E., & Margolis, E.Q. Limits to Ponderosa Pine Regeneration following Large High-Severity Forest Fires in the United States Southwest. *Fire Ecology*. 2018, 14, 143–163.

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*. 382-401.

Johnson, J.B., & Omland, K.S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution*. 19(2), 101-108.

Katuwal, H., Calkin, D.E., & Hand, M.S. (2016). Production and efficiency of large wildland fire suppression effort: a stochastic frontier analysis. *Journal of Environmental Management.* 166, 227-236.

Keane, R.E., Agee, J.K., Fulé, P., Keeley, J.E., Key, C., Kitchen, S.G., Miller, R. & Schulte, L.A. (2009). Ecological effects of large fires on US landscapes: benefit or catastrophe? A. *International Journal of Wildland Fire*. *17*(6), 696-712.

Korobkin, R., & Guthrie, C. (2003). Heuristics and biases at the bargaining table. *Marquette Law Review.* 87, 795.

Larkin, N.K., Raffuse, S.M., & Strand, T.M. (2014). Wildland fire emissions, carbon, and climate: US emissions inventories. *Forest Ecology and Management*. 317, 61-69.

Littell, J.S., McKenzie, D., Peterson, D.L., & Westerling, A.L. (2009). Climate and wildfire area burned in western US ecoprovinces, 1916–2003. *Ecological Applications*. 19(4), 1003-1021.

Littell, J.S., McKenzie, D., Kerns, B.K., Cushman, S., & Shaw, C.G. (2011). Managing uncertainty in climate□driven ecological models to inform adaptation to climate change. *Ecosphere*. 2(9), 1-19.

Liu, Y., Goodrick, S., & Heilman, W. (2014). Wildland fire emissions, carbon, and climate: Wildfire–climate interactions. *Forest Ecology and Management*. 317, 80-96.

Meyn, A., White, P.S., Buhk, C., & Jentsch, A. (2007). Environmental drivers of large, infrequent wildfires: The emerging conceptual model. *Progress in Physical Geography*. 31(3), 287–312.

Moeltner, K., Kim, M.K., Zhu, E., & Yang, W. (2013) Wildfire smoke and health impacts: A closer look at fire attributes and their marginal effects. *Journal of Environmental Economics and Management*. 66, 476–496.

Morgan, M.G., Henrion, M., & Small, M. (1992). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press.

Moritz, M.A., Batllori, E., Bradstock, R.A., Gill, A.M., Handmer, J., Hessburg, P.F., Leonard J, McCaffrey S, Odion, D.C., Schoennael T., & Syphard, A.D. (2014). Learning to coexist with wildfire. *Nature*. 515(7525), 58.

Neary, D.G., Gottfried, G.J., & Ffolliott, P.F. Post-wildfire watershed flood responses. In Proceedings of the 2nd International Fire Ecology Conference, American Meteorological Society, Orlando, FL, USA, 28 November–2 December 2003; Volume 65982.

North, M.P., Stephens, S.L., Collins, B.M., Agee, J.K., Aplet, G., Franklin, J.F., & Fule, P.Z. (2015). Reform forest fire management. *Science*. 349(6254), 1280-1281.

Pelletier, J.D., & Orem, C.A. (2014). How do sediment yields from post□wildfire debris□laden flows depend on terrain slope, soil burn severity class, and drainage basin area? Insights from airborne□LiDAR change detection. *Earth Surface Processes and Landforms*. *39*(13), 1822-1832.

Peppin, D.L., Fulé, P.Z., Sieg, C.H., & Beyers, J.L.; Hunter, M.E.; Robichaud, P.R. Recent trends in post-wildfire seeding in western US forests: Costs and seed mixes. (2011) *International Journal of Wildland Fire*. 20(5), 702–708.

Podschwit, H., & Cullen, A. (2020). Patterns and trends in simultaneous wildfire activity in the United States from 1984 to 2015. *International Journal of Wildland Fire*. 29(12), 1057-1071.

Raftery, A.E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*. 98(464), 931-938.

Reid, C.E., Brauer, M., Johnston, F.H., Jerrett; M, Balmes, J.R., & Elliott, C.T. (2016) Critical review of health impacts of wildfire smoke exposure. *Environmental Health Perspectives*. 124, 1334.

Robichaud, P.R., Rhee, H., & Lewis, S.A. (2014). A synthesis of post-fire Burned Area Reports from 1972 to 2009 for western US Forest Service lands: trends in wildfire characteristics and post-fire stabilisation treatments and expenditures. *International Journal of Wildland Fire*. *23*(7), 929-944.

Rocca, M.E., Miniat, C.F., & Mitchell, R.J. (2014) Introduction to the regional assessments: Climate change, wildfire, and forest ecosystem services in the USA. *Forest Ecology and Management*. 327, 8.

Rodrigues, M., & de la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*. 57, 192-201.

Stavros, E.N., Abatzoglou, J.T., McKenzie, D., & Larkin, N.K. (2014). Regional projections of the likelihood of very large wildland fires under a changing climate in the contiguous Western United States. *Climatic Change*. 126(3-4), 455-468.

Stephens, S.L., Burrows, N., Buyantuyev, A., Gray, R.W., Keane, R.E., Kubian, R., Liu, S., Seijo, F., Shu, L., Tolhurst, K.G. & Van Wagtendonk, J.W. (2014) Temperate and boreal forest

mega-fires: Characteristics and challenges. *Frontiers in Ecology and the Environment.* 12(2), 115–122.

Sugerman, D.E., Keir, J.M., Dee, D.L., Lipman, H., Waterman, S.H., Ginsberg, M., & Fishbein, D.B. (2012). Emergency health risk communication during the 2007 San Diego wildfires: comprehension, compliance, and recall. *Journal of Health Communication*. 17(6), 698-712.

Syphard, A.D., Sheehan, T., Rustigian-Romsos, H., & Ferschweiler, K. (2018). Mapping future fire probability under climate change: Does vegetation matter?. *PloS one*. 13(8), e0201680.

Syphard, A.D., Radeloff, V.C., Keeley, J.E., Hawbaker, T.J., Clayton, M.K., Stewart, S.I., Hammer, R.B. (2007). Human influence on California fire regimes. *Ecological Applications.* 17(5), 1388-1402.

Syphard, A.D., Keeley, J.E., Pfaff, A.H., Ferschweiler, K. (2017). Human presence diminishes the importance of climate in driving fire activity across the United States. *Proceedings of the National Academy of Sciences*. 144(52), 13750-13755.

Taylor, S.W., Woolford, D.G., Dean, C.B., & Martell, D.L. (2013). Wildfire prediction to inform fire management: statistical science challenges. *Statistical Science*, 28(4), 586-615.

Teigen, K.H., & JØrgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(4), 455-475.

Val Martin, M., Heald, C.L., Ford, B., Prenni, A.J., & Wiedinmyer, C. (2013) A decadal satellite analysis of the origins and impacts of smoke in Colorado. *Atmospheric Chemistry and Physics*. 13(15), 7429–7439.

Viegas, D.X., & Simeoni, A. (2011). Eruptive behaviour of forest fires. *Fire Technology*. 47(2), 303-320.

Willmott, C.J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society.* 63(11), 1309-1313.

Zargar, A., Sadiq, R., Naser, B., & Khan, F.I. (2011). A review of drought indices. *Environmental Reviews*, *19*(NA), 333-349.

Zou, Y., O'Neill, S.M., Larkin, N.K., Alvarado, E.C., Solomon, R., Mass, C., Liu, Y., Odman, M.T. & Shen, H. (2019). Machine learning-based integration of high-resolution wildfire smoke simulations and observations for regional health impact assessment. *International Journal of Environmental Research and Public Health*. *16*(12), 2137.

Chapter 2. UNCERTAINTY IN MODEL CHOICE FOR FORECASTS OF VERY-LARGE FIRE OCCURRENCES'[1]

2.1 INTRODUCTION

Although representing only a small fraction of the total number of fires, very-large fires (VLFs) are events often associated with dramatic economic, human health, and environmental risks that are unlike most other wildfires. The most salient and immediate economic impacts are suppression costs and property losses (Barrett 2018), which are often relatively large in VLFs compared to other smaller events (González-Cabán 1983, Stephens et al., 2014). In addition to these direct costs, there is a suite of indirect economic impacts—such as damages from post-fire hazards, rehabilitation costs, lost tax and business revenue from community evacuations (Dale 2009)—that are increasingly probable and costly in larger wildfires (Neary et al., 2003, Peppin et al., 2011, Beverly and Bothwell 2011, Beverly et al., 2011). VLFs have the potential to burn large areas of vegetation and emit tremendous quantities of smoke within a short duration of time, which can adversely impact air quality for months at a time (Stephens et al., 2014), even at long distances from any active burning (Forster et al., 2001, Val Martin et al., 2013). The large areas of active burning and sudden increase in air pollutants pose numerous risks to public health (Reid et al., 2016) and safety (Achtemeier 2009, Stephens et al., 2014); and hospital admissions and treatment costs are expected to increase during VLFs (Moeltner et al., 2013). Although there are some ecological benefits of fire, VLFs have also been associated with significant, deleterious, and

[1] Podschwit, H. R., Larkin, N. K., Steel, E. A., Cullen, A., & Alvarado, E. (2018). Multi-model forecasts of very-large fire occurrences during the end of the 21st century. *Climate*, *6*(4), 100.

sometimes irreversible environmental changes. These include the production of environmental conditions conducive to the establishment of invasive species (Crawford et al., 2001), loss of ecosystem services (Rocca et al., 2014), and long-term modifications to forest structure (Haffey et al., 2018). Given the disproportionate costs VLFs pose to economic, social and environmental values, there is a broad need across disciplines, to better understand patterns and trends of their occurrence to mitigate future hazards. There is even greater urgency for this given that multiple lines of evidence suggest that VLF frequency has increased (Williams 2013, Dennision et al., 2014, Barbero et al., 2014) and will continue to increase into the future (Stavros et al., 2014, Barbero et al., 2015). Still, there are several challenges to obtaining reliable predictions of future wildfire activity, many of which are related to various kinds of scientific unknowns or uncertainties. These uncertainties can come from many sources (Chen et al., 2018) and can be associated with a particular quantity of interest—what will the future frequency of very-large fire events be?—or associated with model structures—how are environmental conditions related to very-large fire frequency? The latter is referred to as model or structural uncertainties, which can have significant effects on the conclusions one draws from an analysis (Morgan et al., 1992, Syphard et al., 2018). In the context of forecasting future wildfire activities, these structural uncertainties can arise from the selection of vegetation models (Sitch et al., 2008, Syphard et al., 2018), assumed anthropogenic effects (Westerling et al., 2011), as well as greenhouse gas emissions and their effects on the environment.

Structural uncertainties associated with the characteristics of the future environment are commonly accounted for in long-term climate impact studies through the use of multiple General Circulation Models (GCMs). Each GCM predicts climatological responses based on unique assumptions regarding chemical and physical interactions between a suite of factors including

land, water, atmosphere, and the cryosphere. These models can be used to forecast future climate by forcing the models to observe historical atmospheric conditions and running the models forward using representative concentration pathways (RCPs) as plausible carbon emission scenarios. There are four future RCP scenarios, which are labeled RCP 8.5, RCP 6, RCP 4.5, and RCP 2.6; each assumes various levels of fossil fuel use and economic activity. The suffix labels correspond to the approximate 2100 radiative forcing levels. For instance, the high-emission (RCP 8.5) scenario corresponds to an approximate radiative forcing increase of 8.5 W/m^2 by 2100 compared to pre-industrial conditions. Since the choice of GCM and RCP can be thought of as competing plausible models of the future environment, and the precise climatological response and quantities of greenhouse gases that will be emitted and sequestered prior to 2100 are unknown, it makes sense to interpret them as structural uncertainties (Taylor et al., 2012).

In addition to the structural uncertainty arising from relating greenhouse gas emission scenarios to climatological impacts, structural uncertainty further arises when relating climatological variables to an impact of interest. In the context of fire, these relationships could include weather patterns such as temperature, precipitation, atmospheric moisture, winds, and clouds. Identifying and describing the relationship between these variables and VLFs is an immensely complex and subjective process, as there can be many competing hypotheses. For instance, temperature controls landscape flammability, is associated with thunderstorm activity and by extension ignition frequency (Flannigan et al., 2009), mediates tree mortality through drought (Allen et al., 2010) and insect pests (Bentz et al., 2010), and influences the length of the snow-free season (Westerling 2016). Additionally, the timing and amount of precipitation can also influence wildfire behavior in parallel with temperature by controlling the availability of fine fuels (Meyn et al., 2007), fuel moisture (Flannigan et al., 2016), and distribution of flammable species

(Bradley et al., 2016). Multiple weather variables may adequately measure a common phenomenon associated with wildfire risk, such as drought (Zargar et al., 2011), resulting in highly correlated covariates that when utilized in wildfire risk prediction, produce models with near-identical goodness-of-fit. Hence, although statistical models can be a useful tool for representing and identifying the relative importance of relationships between the environment and fire, they are still only approximations to reality. Confounding of unmeasured variables like vegetation management (Holsinger et al., 2016) may influence the predicted importance of some weather variables, which can make model selection challenging. In some cases, the same suite of covariates can be used to make predictions with multiple mathematical representations (Mallick and Gelfand 1994), presenting an additional level of uncertainty that is easily overlooked. Given the frequency with which one faces structural uncertainties when modeling highly complex phenomena like wildfire, it is extremely risky to select any single model as an approximation of this phenomena, and a more robust approach should explore results from multiple models (Morgan et al., 1992, Littel et al., 2011).

Bayesian model averaging is flexible and a commonly used method of accounting for structural uncertainties like these, lessening many of the risks of traditional model selection techniques and improving performance across a variety of metrics (Raftery and Zheng 2003). Within this framework, model weights, which are assumed to be uncertain, are used to combine covariates or predictions from multiple sources into a single probability distribution. The uncertainties in the model weights are represented using the posterior, which is a probability distribution representing the belief in the model parameters conditional on the observed data. While posteriors provide a natural framework for interpreting uncertainties, in practice, closed form expressions of the posterior are non-trivial and direct calculation is often impossible. Hence,

simulation methods like Markov Chain Monte Carlo are typically used to generate samples from the distribution, which are in turn used to approximate the quantities that are of interest to the analyst (Fragoso et al., 2018). Computational barriers to Markov Chain Monte Carlo techniques have diminished greatly since they were first introduced, and a range of recently developed software options such as JAGS (Plummer 2003), Stan (Carpenter et al., 2017), and Integrated Nested Laplace Approximations (Rue et al., 2009) have facilitated the application of these methods in novel and previously infeasible contexts (Monnahan et al., 2017).

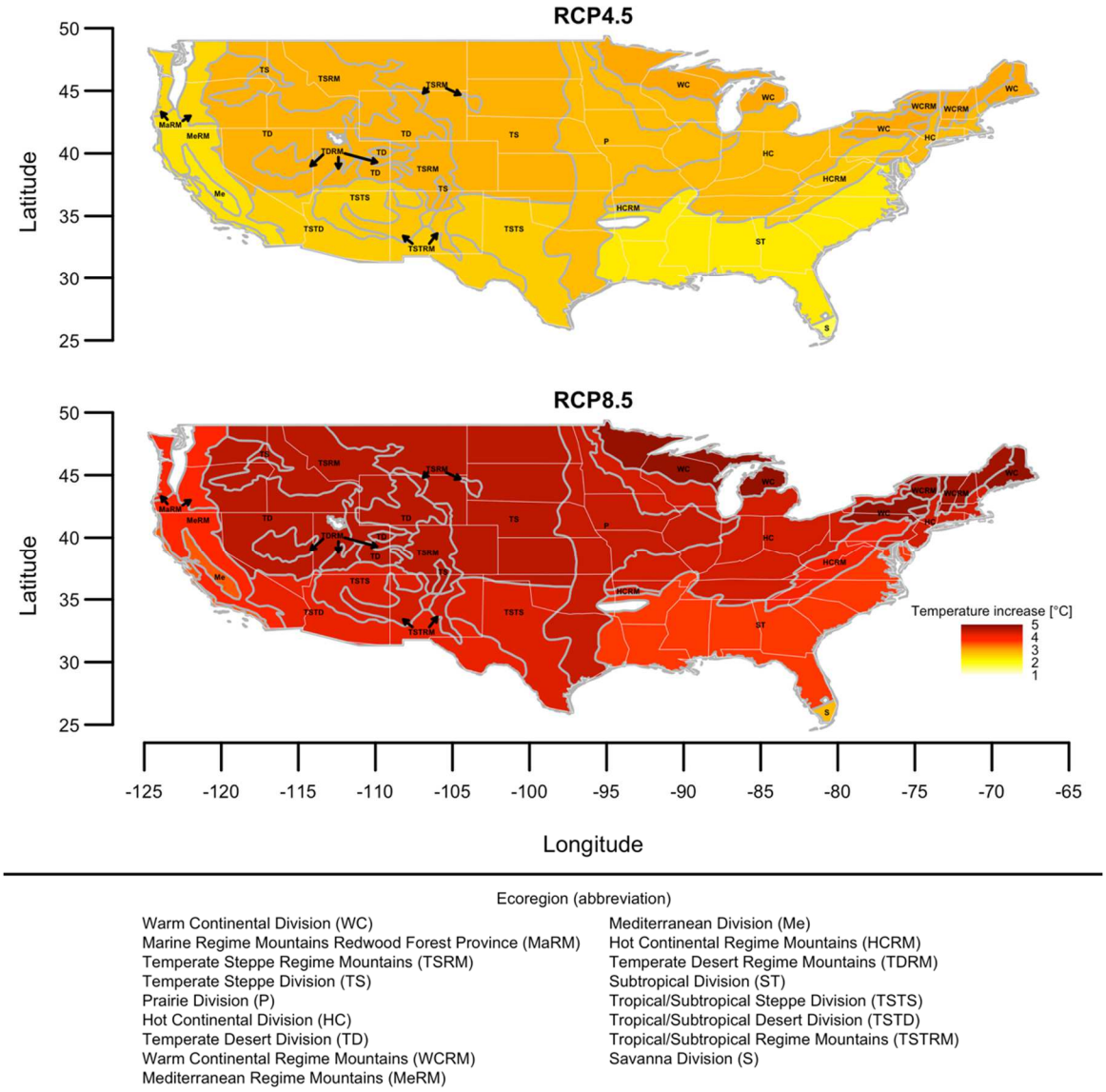
Hence, in this paper, I account for both kinds of structural uncertainty—uncertainty from the climate models and uncertainty from the choice of VLF models—using Bayesian model averaging to generate predictions of event frequency in the last half of the 21st century in the Continental United States. In Section 2, I present the methods used to produce robust predictions of future wildfire activity using GCMs and multiple fire occurrence models. In Section 3, the results of this analysis are available and demonstrate that increases in VLF activity should be expected in many regions in the Continental United States at the end of the century. In Section 4, I close the paper with a discussion of the implications of this analysis to decision-makers and researchers.

2.2 METHODS

2.2.1 *Fire Occurrence Data*

Data from the Monitoring Trends in Burn Severity (MTBS) project, which describes individual fire size and severity based on changes in satellite imagery, are used to measure monthly fire occurrence. The original dataset included all detected fire events within the continental United States for the years 1984–2015 and is further filtered to remove all events 404 hectares or smaller, or that were non-wildfire. The filtered data are grouped into 18 regions with broadly

similar climate and vegetation characteristics using a geospatial dataset of ecosystem divisions (Figure 2.1, Bailey 2016). For each region, two binary time series were constructed: one representing large fire (LF) occurrence, and another representing very-large fire (VLF) occurrence. The LF occurrence time series, $X_{LF,t}$ reports "1" if at least 1 event of at least 404 hectares is recorded during that month; otherwise, it reports "0". The VLF occurrence time series, $X_{VLF,t}$, records "1" if at least 1 VLF—one that exceeds the 95th percentile of the region's filtered MTBS burn area records (Table 2.1)—is recorded during that month and region; otherwise, it reports "0". The Marine Division had one fire event and the Subtropical Regime Mountain had no VLF events during 1984–2005, and both were dropped from further consideration, yielding a total of 16 independent ecoregional analyses. All of the time series are split into a tuning dataset (1984–2005) and a training dataset (2006–2015).



Ecoregion (abbreviation)

Warm Continental Division (WC)
Marine Regime Mountains Redwood Forest Province (MaRM)
Temperate Steppe Regime Mountains (TSRM)
Temperate Steppe Division (TS)
Prairie Division (P)
Hot Continental Division (HC)
Temperate Desert Division (TD)
Warm Continental Regime Mountains (WCRM)
Mediterranean Regime Mountains (MeRM)
Mediterranean Division (Me)
Hot Continental Regime Mountains (HCRM)
Temperate Desert Regime Mountains (TDRM)
Subtropical Division (ST)
Tropical/Subtropical Steppe Division (TSTS)
Tropical/Subtropical Desert Division (TSTD)
Tropical/Subtropical Regime Mountains (TSTRM)
Savanna Division (S)

Figure 2.1. Map of multi-model mean temperature changes between 1955–2005 and 2050–2099 over the relevant Bailey's divisions under the representative concentration pathway (RCP) 4.5 (top) and RCP 8.5 emission scenarios (bottom). Regions with insufficient fire occurrence data for analysis are colored white.

Table 2.2. Very-large fire cutoffs, the number of fires, large fire months, and very-large fire months for two time periods: 1984–2005 and 2006–2015.

DOMAIN	Division	95th size percentile (ha)	Total # of fires		# of large fire months		# of very-large fire months	
			'84-'05	'06-'15	'84-'05	'06-'15	'84-'05	'06-'15
DRY	Temperate Desert (TD)	16,007	1623	833	119	55	22	19
	Temperate Desert Regime Mountains (TDRM)	11,765	121	78	52	31	5	4
	Temperate Steppe (TS)	11,664	464	347	119	70	13	13
	Temperate Steppe Regime Mountains (TSRM)	19,853	798	541	101	56	14	13
	Tropical/Subtropical Desert (TSTD)	11,616	365	254	94	57	10	9
	Tropical/Subtropical Regime Mountains (TSTRM)	13,169	168	143	71	45	7	7
	Tropical/Subtropical Steppe (TSTS)	10,243	388	546	126	79	10	16
Temperate	Hot Continental (HC)	6180	267	75	67	41	11	2
	Hot Continental Regime Mountains (HCRM)	4586	169	45	30	27	4	1
	Marine Regime Mountains Redwood Province (MaRM)	21,776	136	130	48	33	6	5
	Mediterranean (Me)	10,980	149	64	82	35	6	3
	Mediterranean Regime Mountains (MeRM)	17,772	799	374	143	60	16	17
	Prairie (P)	6707	155	275	47	56	5	9
	Subtropical (ST)	5908	431	312	149	82	16	14
	Subtropical Regime Mountains (SRM)	3927	4	16	3	12	0	1
	Warm Continental (WC)	6466	73	20	40	12	1	4
Humid	Savanna (S)	20,623	89	45	44	24	5	2

2.2.2 *Meteorological covariates*

Regional averages of gridded weather variables from the University of Idaho gridMET dataset are used to calculate 12 weather predictors that will provide coarse scale environmental descriptions during each month between 1984–2015. Of these 12 weather predictors, four are measures of temperature; six are measures of moisture levels, and two measure wind characteristics. The four temperature metrics are based on monthly space-time averages of daily

average temperature, which are calculated by dividing the sum of the daily maximum and minimum values by two (Weiss et al., 2005). The quantity hereafter referred to as seasonality measures intra-annual temperature variability by normalizing monthly temperature averages by the mean and standard deviation of all 360 measurements in the most recent 30 years of data (e.g., 1986–2015). The inter-annual temperature variability is captured with a quantity referred to as the departure from normal, which instead normalizes by the mean and standard deviation of 30 measurements in the most recent 30 years of data that correspond to same month as the raw measurement. The remaining temperature metrics are the rolling 12-month minimum and maximum temperature, which will record extreme temperature events that have potential for delayed impacts on wildfire activity. The six moisture level metrics are average specific humidity and precipitation totals over five time periods (1, 3, 6, 12 and 24-month time windows). In addition to a simple space-time average of wind speed at 10 m, the maximum daily space-time averages each month was also included as a covariate of the fire occurrence probabilities.

2.2.3 *Probability estimation trees*

Two quantities are estimated for each month and region, the probability that at least one LF (>404 hectares) occurs and the probability that a VLF occurs conditional on the occurrence of at least one LF. These probabilities are estimated using multi-model averages of a flexible and powerful type of binary classifier known as a probability estimation tree (PET). PETs use decision-tree structures to recursively divide the data with binary splits, eventually grouping all the data into mutually exclusive categories or leaves. With respect to the response, the splits create increasingly homogeneous clusters of observations, which also occupy an increasingly specific portion of the covariate space. Within the context of this analysis, I have 12

meteorological predictors available to form these categories, so that months—in which certain fire events did or did not occur—can be grouped into categories describing broadly similar environmental conditions. Prediction is performed by using the relevant covariates to identify the appropriate category and taking the empirical frequency of the binary responses in that category as a probability estimate (Provost and Domingos 2000). While it is well known that predictions based on individual decision tree algorithms can be highly variable with significant levels of structural instability (Wang et al., 2016), these pathologies are often lessened through the use of model averaging (Provost and Domingos 2000). To that end, a suite of 100 PETs are generated for each region and for both probabilities of interest; and I will hereafter refer to each collection of 100 PETs as a 'forest'. Each individual PET within a forest is generated stochastically by applying the C4.5 learning algorithm without pruning (Quinlan 1993; Provost and Domingos 2000) to a random sample of the training dataset via the Roughly Balanced Bootstrapping algorithm (Hido et al., 2009). The LF forests and the conditional VLF forests are constructed somewhat differently in that the LF forests sample from all months in the training dataset, while the VLF forests are based only on samples of months in which at least one LF has occurred. In other words, the LF forests will discriminate between LF and no-fire months, and the VLF forests will discriminate between LF and VLF months. Identification of important predictors within each forest are assessed using two summary statistics: (1) the frequency that a predictor is present in the PETs; and (2) the frequency that a predictor is used in the first split of the PETs. The former identifies the frequency with which a given meteorological predictor is used at all within the forest, and the latter identifies the frequency with which a meteorological predictor is the best determinant of the response on a randomly generated dataset.

2.2.4 *Multi-model very-large fire predictions*

The structural uncertainties arising from training PETs to observed meteorological data are compounded by structural uncertainties arising from the application of these models to long-term climate forecasts. To improve the quality of the probability estimates and VLF occurrence forecasts, multi-model averages of fire event probabilities are used to integrate both sources of structural uncertainty: the choice of the PET and selection of the climate model. The final probability estimates are assumed to be an average of predictions from all combination pairs of the 100 PETs within each forest and 13 modeled weather datasets; a total of 1300 individual predictions for each region, month, and probability of interest. The modeled weather data used to make PET predictions come from the second version of regional Multivariate Adaptive Constructed Analogs (MACA) dataset that was trained on gridMET, and downscaled with 13 GCMs: bcc-csm1-1-m, BNU-ESM, CanESM2, CCSM4, CNRM-CM5, CSIRO-Mk3-6-0, GFDL-ESM2M, HadGEM2-ES365, inmcm4, IPSL-CM5A-MR, MIROC5, MRI-CGCM3, NorESM1-M.

The multi-model averages are calculated using both unweighted and weighted approaches. The unweighted approach assigns equal weight to each individual prediction and assumes that each PET and climate model is equally credible. The weighted approach assigns unequal weights to predictions from each pairing of PETs and climate models, to try to bias-correct the multi-model averages and optimize predictive performance. The final weight applied to each individual prediction is the product of two independent components: a climate model weight and a PET weight. For a specified region and month, the estimated probabilities are written as (Equations 2.1-2.2),

$$\bar{p}_{LF} = \sum_{i=1}^{100} \sum_{j=1}^{13} u_{LF,i} v_j p_{LF,i,j}, \qquad (2.1)$$

$$\bar{p}_{VLF} = \sum_{i=1}^{100} \sum_{j=1}^{13} u_{VLF,i} v_j p_{VLF,i,j}. \qquad (2.2)$$

Here, $u_{*,i}$ represents the weight applied to the predictions from PET i, v_j represents the weight applied to predictions utilizing climate model j, and $p_{*,i,j}$ is the prediction obtained from PET i utilizing climate model j. I estimate the weight components using a fully Bayesian approach that incorporates fire occurrence and modeled climate forcings from 1984–2005, as well as probabilistic representations of possible parameter estimates. Via Bayes rule, the posterior of the model weight components, $\theta = \vec{u}_{LF}, \vec{u}_{VLF}, \vec{v}$, is proportional to the product of the likelihood and prior probability distributions. The likelihood component represents the probability of observing the fire occurrence time series, $X = \vec{X}_{LF}, \vec{X}_{VLF}$, assuming that they were generated from a Bernoulli process parameterized with our weighted multi-model averages (Equations 2.3-2.4):

$$\vec{X}_{LF} \sim \text{Bernoulli}(\bar{p}_{I,F}), \quad (2.3)$$

$$\vec{X}_{VLF} \sim \text{Bernoulli}(\bar{p}_{VLF}), \quad (2.4)$$

The prior component, $p(\theta)$, which is a probability density function representing our a priori belief regarding the parameter values, is defined using independent Dirichlet priors with uninformative concentration parameters (Equation 2.5):

$$\vec{u}_{LF}, \vec{u}_{VLF}, \vec{v} \sim \text{Dirichlet}(\vec{1}). \quad (2.5)$$

The posterior was approximated using Just Another Gibbs Sampler (JAGS) software (Plummer 2003) and the runjags package in R (R Development Core Team 2008). An initial set of 30,000 samples were generated from three parallel Markov Chain Monte Carlo chains using a burn-in interval of 10,000 steps, adaptive phase of 10,000 steps, and thinning interval of 100. Calculations were performed on a MacBook Pro (Quanta Computer, Inc, Shanghai, China) with a 2.7 GHz Intel Core i7 processor (Hillsboro, Oregon, USA). Convergence was monitored visually, and also via the calculation of the potential scale reduction factor, using the range of the

central 90th percentile of the marginal posteriors as a test statistic (Brooks and Gelman 1998). The second half of the chain is assumed to have approximately converged if the maximum potential scale reduction factor fell below 1.01. If the chain had not converged, then it was continued in batches of 1000 iterations until the maximum potential scale reduction factor was less than 1.01 (Gelman and Shirley 2011). To provide guidance to future analysts looking to perform similar analyses, an informal computational comparison of JAGS (Version 4.3.0) and Stan software (Version 2.17.3) was completed, which is described in the supplementary materials. The final VLF probabilities can then be calculated by averaging both probabilities of interest—either with point estimate averages of the model weights or using the full posterior in the weighted approach—and applying Bayes rule (Figure 2.2). The resulting distribution of VLF probability time series is then used to estimate the changes in VLF frequency in the future by finding the difference between the expected number of VLFs in the historical climate (1956–2005), and the expected number of VLFs in the future (2050–2099) under moderate and severe warming scenarios, RCP 4.5 and RCP 8.5, respectively.

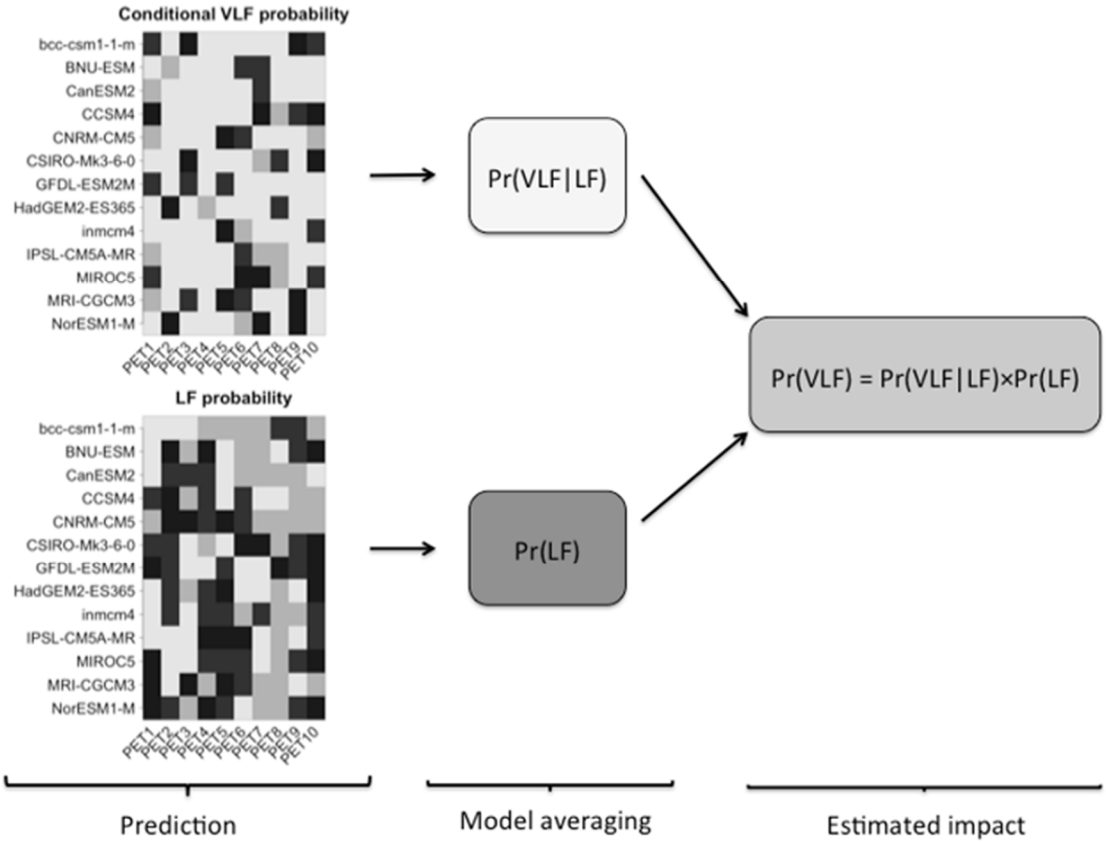


Figure 2.2. Workflow of very-large fire probability calculations. In the prediction phase, probability estimates are calculated for every combination pair of climate model and probability estimation tree. The model-averaging phase combines these predictions into probability estimates using either a weighted or unweighted averages. The final phase uses Bayes rule to calculate the very-large fire probability as the product of both components from the model-averaging phase.

2.2.5 *Ensemble assessment*

The ability of the multi-model averages to estimate observed VLF frequencies was quantified over the temporal range of the extant MTBS fire occurrence record at three non-overlapping time periods. Two of the three time periods correspond to the tuning (1984–2005), training (2006–2015) datasets that were used to bias correct and fit the initial suite of PETs respectively. Additionally, a testing dataset independent of the information used to build the multi-model averages was constructed using 2016 MTBS occurrence data. For each time period, a sample of 100,000 probability time series were drawn from the relevant multi-model average posterior,

which were then used to simulate the distribution of VLF counts predicted during that time period. Note that the 2016 fire data used to independently validate the multi-model averages represent an updated version of the MTBS data that was unavailable during the PET training and tuning stages, and that slight differences in the total number of large (>404 hectares) incidents between 1984–2015 were observed in the two versions. Specifically, the original MTBS dataset reported 10,295 large incidents between 1984–2015, while the updated version reported 10,298 large incidents during that same period.

2.3 RESULTS

2.3.1 *Important predictors of very-large fires*

The diversity of predictors used in the PETs was high, and the important meteorological variables varied by region, the summary statistic, and the type of fire probability. Temperature metrics, in particular seasonality, are a commonly utilized weather predictor in LF forests, and in 10 of the 16 LF forests, seasonality is present in 90 or more of the PETs. On the other hand, while temperature metrics are frequently utilized when constructing PETs, they are not always the optimal splitting criterion. For instance, in the Savanna, Prairie, and Hot Continental Regime Mountains, precipitation metrics overwhelmingly replace temperature-based metrics as the optimal discriminant of large and no fire months, and in other regions such as the Subtropical and Hot Continental Division, this designation is highly uncertain.

The importance of temperature metrics also varied by the type of fire probability considered, with temperature metrics more commonly identified as the optimal split criterion in LF forests compared to VLF forests. This sensitivity of PET structure to the type of fire probability could also arise in other ways. For example, in the Hot Continental Regime Mountains, the LF forest

overwhelmingly relies on precipitation metrics for prediction, while the corresponding VLF forest utilizes no predictors and reports a constant conditional VLF probability. Similarly, in the Tropical/Subtropical Regime Mountains and Prairie divisions, conditional VLF forests tended to identify wind metrics as the optimal split criterion much more frequently than in the LF forests. Additionally, PET complexity tended to be lower in VLF forests than in the LF forests. The average number of variables used per PET, size, and the number of leaves were inflated in the latter, and weather invariant null models were only ever observed in the VLF forests. The variability in the optimal splitting criterion was also higher in the VLF forests, suggesting a relative lack of certainty regarding the optimal discriminant in conditional VLF probabilities compared to LF probabilities. Weighting did not appear to drastically influence the relative contribution of the weather predictors within each forest (Figure 2.3).

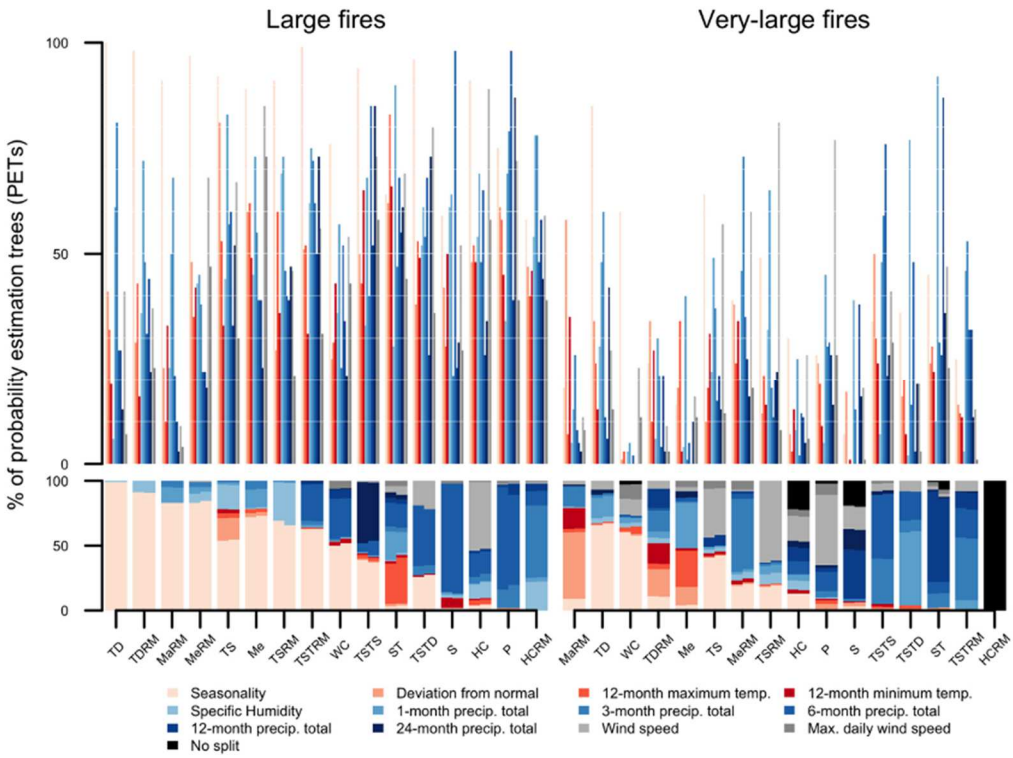


Figure 2.3. Summary statistics of each region's forest of probability estimation trees. The top panels show the percentage of probability estimation trees (PETs) in each region's forest that uses a particular weather predictor for at least one split. The bottom panels show the percentage of PETs for which a weather predictor is selected as the optimal splitting criterion. The regions are listed along the x-axis. The relative contribution of each first-split variable under the unweighted (left) and weighted (right) averaging methods are displayed side-by-side. No split refers to PETs that predict constant probabilities in all meteorological conditions.

2.3.2 *Climate change and very-large fire occurrence*

In most divisions, the expected number of VLFs is predicted to increase in 2050–2099 compared to 1956–2005. The Marine Regime Mountains Redwood Forest division is predicted to have the largest absolute increase with about 13 additional fires per decade under the RCP 4.5 scenario and about 18 additional fires per decade under the RCP 8.5 scenario. For most of the regions under consideration, the average predicted increase ranges from near-zero to several additional VLF per decade relative to historical predictions. In some regions, like Mediterranean and Savanna divisions, the multi-model average predicts slight decreases in VLF activity. The largest

absolute decrease occurred in Mediterranean California, which is predicted to have about one less VLF per decade relative to historical predictions under both RCP scenarios. In general, increases in VLF frequency are more severe under the RCP 8.5 scenario than under the RCP 4.5 scenario, although the sensitivity to RCP scenario varies by division (Figure 2.4). The largest absolute difference in average VLFs per decade between the RCP 8.5 and RCP 4.5 scenarios was in the Marine Regime Redwood Forest division, which had about 5 additional VLFs in the RCP 8.5 scenario. Although the Hot Continental Regime Mountains predicts a larger VLF count per decade under the RCP 4.5 scenario than the RCP 8.5, the difference is negligibly small. The median difference between the RCP 8.5 and RCP 4.5 scenario across the 16 ecoregions was 0.9 additional VLFs per decade under the RCP 8.5 scenario.

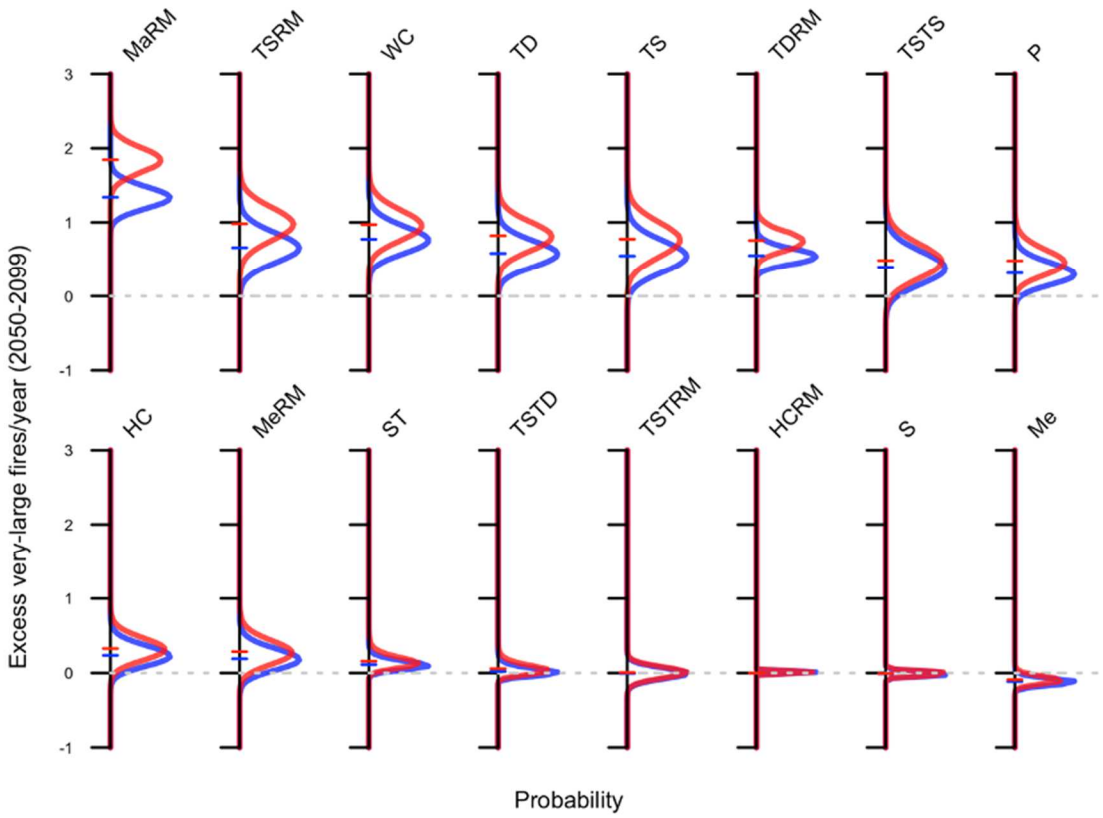


Figure 2.4. Kernel density estimates of the posterior and mean of the number of additional very-large fires per year relative to the 1956–2005 reference period per year by ecoregion under the representative concentration pathway (RCP) 4.5 (blue) and RCP 8.5 (red) scenarios arranged by magnitude of change. The excess very-large fire frequency is calculated by randomly sampling ($n=10^6$) from the posterior of historical (1956–2005) and future (2050–2099) multi-model averages and calculating the difference.

Future changes in VLF frequency may or may not be uniformly distributed throughout the year. The largest absolute monthly changes in VLF frequency are observed in the Marine Regime Mountain Redwood Forest division during the summer months, while the shoulder months are not predicted to drastically differ from present day VLF frequency. In contrast to the predictions in the Marine Regime Mountain Redwood Forest division, semi-uniform changes in VLF frequency are also predicted in some regions. For instance, the Subtropical division is predicted to have about 6-8 additional VLF events during the last half of the 21st century relative to the 1956–2005 reference period but shows no strong preference as to what month these events will

occur. For nearly all regions and months, VLF frequency is predicted to increase or show no change compared to historical reference conditions, with the Prairie division being an example of the former and the Hot Continental Regime Mountains the latter. The Mediterranean division is an exception to this pattern, as reductions in future very-large fire frequency are predicted from October to May (Figure 2.5).

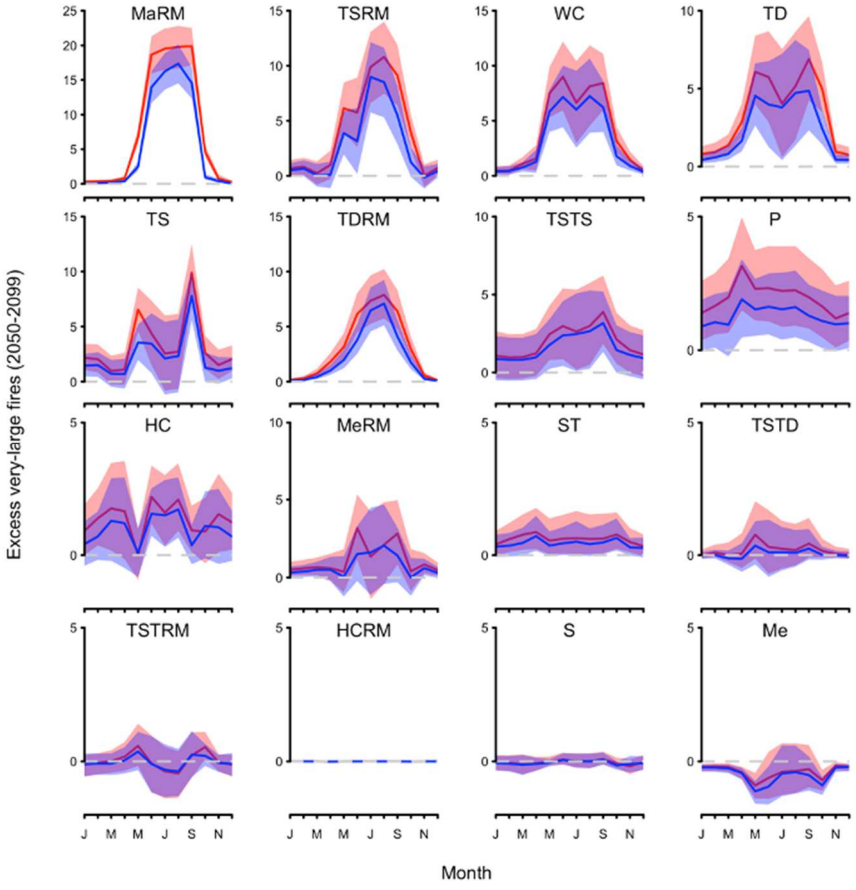


Figure 2.5. Predicted intra-annual changes in very-large fire frequency across sixteen biogeographical regions within the Continental United States under the RCP 4.5 (blue) and RCP 8.5 (red). The central 90th percentile and mean of the excess very-large fires are based on 1,000,000 random samples of the posterior multi-model average very-large fire probabilities from the historical (1956–2005) and future (2050–2099) scenarios.

The changes in overall VLF frequency are predicted to be a result of changes in both model components: the LF and conditional VLF occurrence probabilities. For divisions like Marine

Regime Mountains Redwood Forest, both probabilities increase, implying that the LF months will become increasingly frequent and a larger proportion of the months classified as LF will become VLF months. Other divisions showed increases in only one of the model components. In the Temperate Steppe Regime Mountain division, only conditional VLF probabilities are predicted to increase, and in the Tropical/Subtropical Steppe division, only LF probabilities are anticipated to increase. Significant decreases in the model components are only predicted in the Mediterranean and Tropical Subtropical Regime Mountains divisions, which respectively have decreases in the conditional VLF and LF probability components in 2050–2099 compared to 1956–2005 climate model forcings. The Mediterranean LF probability components are predicted to increase, while the conditional VLF probability component is expected to remain the same in the Tropical Subtropical Regime Mountains division. In general, the changes in model components are greater in the RCP 8.5 scenario compared to the RCP 4.5 scenario, although the differences between the two future scenarios were nearly imperceptible in some regions (Figure 2.6).

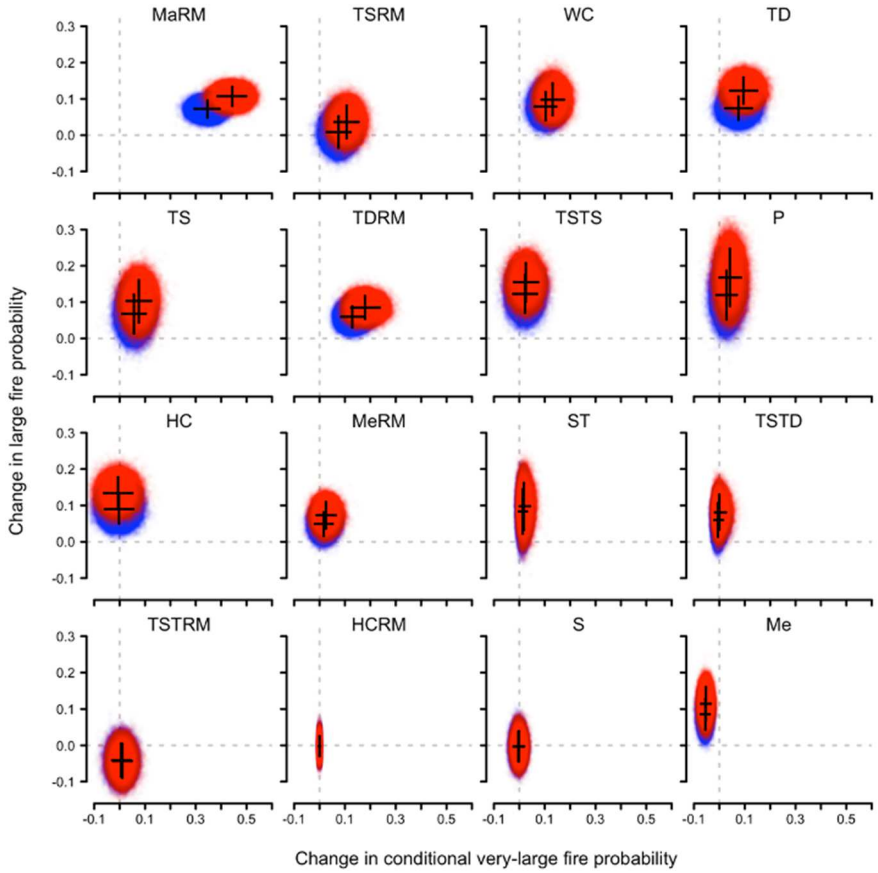


Figure 2.6. Simulated change in monthly multi-model large-fire and conditional very-large fire probability estimates across biogeographical divisions of the continental United States. The point cloud is a sample of 100,000 differences in average posterior probability components under the historical (1956–2005) and future scenarios (2050–2099); with the RCP 8.5 scenario colored red and RCP 4.5 colored blue. The solid black lines represent the central 90th percentile and the dashed lines are horizontal and vertical lines passing through the origin.

2.3.3 *Ensemble assessment*

The proportion of simulated VLF counts equal to or below the observed values varied by region and time period, and the frequency with which this quantity fell within the central 95th percentile of the simulated VLF counts informs us of the overall quality of the multi-model average forecasts. Using this performance metric, the highest ensemble quality occurs in regions where the central 95th percentile of simulated VLF counts covers the observed VLF counts in all three time periods, which was observed in the Marine Regime Mountains Redwood Forest, Prairie, Hot

Continental, Temperate Desert Regime Mountains, and Savanna divisions. In as many regions, this quantity fell in the central 95th percentile for the testing and tuning time periods only, or in the testing and training time periods only. This was observed in the Temperate Steppe Regime Mountains, Temperate Steppe, Temperate Desert, Mediterranean Regime Mountains, and Tropical/Subtropical Steppe divisions. Predictive performance was occasionally poor in the tuning and training time periods, but good during the testing time period, as was observed in the Warm Continental, Subtropical, and Tropical/Subtropical Desert divisions. In the Mediterranean division, the ensembles performed well on the tuning and training time periods but showed poor performance when predicting data they were not already optimized on. The lowest model quality was seen in the Tropical/Subtropical Regime Mountains, where observed VLF counts were covered by the central 95th percentile in the tuning time period only, and in the Hot Continental Regime Mountains, where the central 95th percentile of the simulated VLF counts never covered the observed quantity.

Consistent underestimation, where the observed VLF count was equal to or greater than the median simulated VLF count in all three time periods, was reported in nine of the sixteen regions considered. The magnitude of these underestimates ranged from very minor, as in the Temperate Desert, to quite severe, as in the Hot Continental Regime Mountains. Consistent overestimates were much less frequently observed, with only the Marine Regime Mountains Redwood Forest and Warm Continental divisions reporting observed VLF counts equal to or less than the median simulated VLF counts in every time period. Five regions had VLF counts that were located to the left or right of the median depending on the time period considered. The Temperate Steppe, Prairie, and Tropical/Subtropical Steppe divisions simulations tended to underestimate the reported VLF

counts, while the opposite was observed in Temperate Steppe Regime Mountains and Hot Continental divisions.

The simulated distributions did not appear to be strongly sensitive to the choice of RCP scenario during the temporal extent of the training (2006–2015) and testing (2016) time periods, as there are only slight differences between them during those times (Figure 2.7).

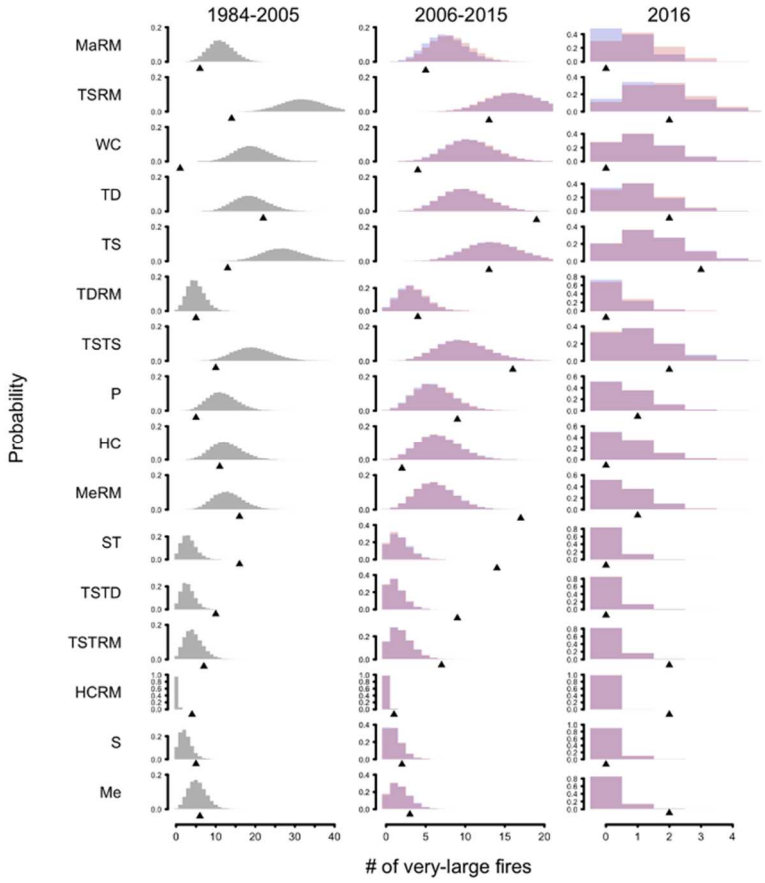


Figure 2.7. Simulated and actual very-large fire month counts for each region and three time periods: 1984–2005, 2006–2015, and 2016. A sample of 100,000 simulated very-large fire (VLF) counts are produced under historical (grey), RCP 4.5 (blue), and RCP 8.5 (red) scenarios by randomly selecting a VLF probability time series from the posterior and randomly generating a VLF occurrence time series. The observed VLF counts are represented with arrows.

2.4 DISCUSSION

2.4.1 *Important predictors of very-large fires*

Wildfire events are associated with a number of factors (Flannigan et al., 2009) that may vary in space (Stavros et al., 2014, Barbero et al., 2015, Arpaci et al., 2013, Flannigan et al., 2006, Brotak and Reifsnyder, W. E. 1977), and may reveal themselves only under certain conditions (Slocum et al., 2010, Krueger et al., 2015); it should not then be unexpected that model variability can often be high. Attempts to identify any single factor as most closely associated with VLFs are frustrated by the complex behavior of wildfires, competition among models, data limitations, and diversity of performance criterion. Despite the ubiquity of structural and other uncertainties, the relative importance of various coarse scale meteorological factors to specific wildfire activities could be gauged by observing the frequency with which they were utilized to make predictions. In some cases, a meteorological variable could, with high confidence, be readily identified as important to predicting VLFs in a particular region. In the Temperate Desert division, seasonality was frequently utilized in PETs for both wildfire probabilities and was also often identified as the optimal splitting criterion. More typically, however, some level of structural uncertainty was present and identifying a best predictor was not always as obvious. In the Subtropical division, LF forest, temperature and precipitation variables were identified as the optimal splitting criterion with nearly equal frequency. In the Mediterranean Regime Mountains division, seasonality was frequently the optimal splitting criterion in the LF forest, but it was much less common in the VLF forest. Moreover, in the Mediterranean division, wind-based metrics were frequently utilized in LF forests in the Mediterranean forest, but not as a first-split in the PETs. Model variability could be particularly high in the LF forests in the Eastern Continental United States. Precipitation based variables were overwhelmingly preferred in

extreme southern Florida and in the Appalachians, but wind-based variables were preferred in the Hot Continental division; Temperature was slightly preferred in the Warm Continental division, and as already mentioned, the Subtropical region showed no strong preference with regard splitting criterion. Although some regions showed preferences for certain weather variables, model variability was fairly high in the VLF forests.

Although these structural uncertainties are sometimes obstacles to identifying important meteorological relationships with VLFs, they are also critical to understanding the true level of confidence in observed correlations and safeguard against overconfident conclusions. While clearly notable levels of model variability could be encountered across multiple factors, robust patterns and trends could still be inferred. For instance, note that, in most of the West, with the exception of the Great Plains and the Tropical/Subtropical portions of the Southwest, temperature metrics were often the best predictor of LFs and were commonly used in LF forests. In the remaining Western areas, temperature metrics were less useful and instead precipitation metrics were selected as the optimal splitting criterion. This apparent preference for precipitation metrics over temperature metrics in these regions may be related to the characteristics of fuel-limited versus climate-limited fire regimes (Meyn et al., 2007), or due to a relative inability of seasonal temperature fluctuations to match wildfire activity compared to precipitation. The relative popularity of wind-based variables in very-large forests compared to very-large forests is also interesting, as wind has been reported to have variable influence on wildfire activity depending on fire size and geographic location (Slocum et al., 2010).

2.4.2 *Climate change and very-large fire occurrence*

For both RCP scenarios and nearly all divisions, complex changes to wildfire activity are predicted that will result in an overall increase in the frequency of VLFs, which is largely consistent with many other projections (Flannigan et al., 2009, Barbero et al., 2015). While overall increases in the frequency of these events are predicted using robust methods, the exact nature of these changes remain unclear. It is not certain, for instance, if the range of fire sizes will remain largely static in the future and only frequency of exceptionally large events will increase; or if the size distribution will shift, so that burn areas exceed historic records. These distinctions are important because the relative costs of these two competing possibilities are likely to vary across decision makers. Put another way, although it would be equivalent within the modeling framework described here, the consequences of adding five 100,000-acre events to the historical wildfire size distribution and of adding five 1,000,000-acre events to the historical wildfire size distribution are not the same to decision makers. The Mediterranean division was somewhat of an exception to the overall reported increases in VLF activity. Westerling et al. (2011) project either no change or modest increases in LF activity in much of lowland California, and large increases in mid- and high-elevation locations, which at first seems inconsistent with the predicted decrease in VLF activity, although there are a few explanations. Firstly, by considering a larger number of climate models and predictive models, the range of results in this analysis will be inherently more variable, and marginal results seen in other studies could emerge as significant when these structural uncertainties are incorporated. Secondly, as shown in this study, the environmental drivers of large and conditional VLF probabilities can vary, and differences regarding the definition of LF can result in variability amongst methodologies (Slocum et al., 2010). Thirdly, differences between the covariates considered and model structure are likely to alter the predictions across analyses.

For instance, anthropogenic and vegetation effects on wildfire activity were omitted in this study but are known to be an important influence of wildfire activity in California (Syphard et al., 2007) and elsewhere (Syphard et al., 2017).

The months in which VLF activity was historically highest may not necessarily apply in the second half of the 21st century, and noticeable changes in intra-annual patterns, usually increases, of VLF activity were predicted in most scenarios and regions. Some regions, like Temperate Desert Regime Mountains and Marine Regime Mountains Redwood Forests, are predicted to have increases in VLF frequency only during a limited portion of the year, while others, like the Subtropical division, are predicted to have a relatively uniform increase in VLF frequency throughout the year. Given that simultaneous increases in VLF probabilities are anticipated in multiple independent regions, it is likely that VLF activity will change in ways that will increase resource strain, which is a result consistent with Podschwit and Cullen (2020). Indeed, the results of this study suggest that, depending on the emission scenario, between 12–13 regions will have future VLF frequencies that exceed the historical record, and that intra-annual increases in VLF occurrence are often predicted during the same time of year in spatially distinct regions.

In addition to changes to intra-annual patterns of overall VLF frequency, it is important to acknowledge that the overall increases in VLF frequency are the product of two processes: changes in LF and conditional VLF probabilities. Any increase in VLF frequency is then the result of one of three scenarios: an increase in both probabilities, and increase in LF frequency only, or an increase in the frequency that LFs become VLFs. These specific changes in model components may be of particular relevance to firefighting, public health professionals, and other decision-makers who will—due to differences in the impacts of the events—react to no-fire, LF, and VLF months differently and require guidance regarding the characteristics of the novel future wildfire

regimes. Reducing the uncertainty as to which emission scenario the future will resemble should also be a priority for decision-makers and researchers, as the predicted changes tend to be more exaggerated under the RCP 8.5 scenario compared to the RCP 4.5, which should influence adaptation and mitigation efforts of future wildfire impacts.

Given the already strained firefighting operations under current wildfire size distributions, the results of this chapter suggest that these suppression strategies are unlikely to be sustainable by the end of the century. Firefighting operations will eventually likely be forced to either (1) increase the funding and resources for firefighting or (2) use less aggressive firefighting strategies that allow some wildfires to grow with little to no suppression. That wildfire activity is higher under the more extreme warming scenario also suggest that these suppression conflicts are likely to be smaller if aggressive carbon mitigation policies are enacted.

2.4.3 *Caveats and future work*

While the simultaneous acknowledgement of structural uncertainties in the climate models and PETs represents an interesting approach, there are still a number of uncertainties that were not addressed in this climate impact analysis. The limited availability of reliable and consistently recorded (e.g., satellite-based) measurements of wildfire activity (Taylor et al., 2013) and the inherent rarity of VLF events remain significant obstacles to validating predictions and estimating underlying model structures (Podschwit and Cullen 2020). The validation results should be considered as the current state of knowledge regarding the ensemble's predictive ability and may change when more data becomes available in the future. If inter-annual variability in wildfire activity is high, then the validation results used in this study may be based on particularly predictable or unpredictable fire years, and therefore not be representative of the actual

performance. Longer duration datasets would be preferred, and thirty year climatologies are often considered ideal (Arguez and Vose 2011), but the entire range of available burn area data only extends 33 years and it is unlikely that longer time scale meteorological associations with VLF activity will be accurately captured with the relative brevity of data (Westerling and Swetnam 2003, Marlon et al., 2012). Moreover, if recent increases in VLF activity are indicative of a sudden a shift into overall wildfire patterns unlike what has been observed in the past, then forecasting future activities based on historical relationships could be inadequate. For instance, the two events occurring in the Hot Continental Regime Mountains in 2016 were quite unusual in historical terms, as only four VLF months were reported from 1984–2005, and only one VLF was reported from 2006–2015.

Data limitations may also be qualitative, and many of the remaining important structural uncertainties are due to unconsidered covariates, like vegetation changes, suppression effort, and population growth, that were not modeled due to data inavailability, practical considerations, and challenges related to predicting these quantities in the future. While the PETs used in this study produced a diverse suite of predictive models and are known to be highly unstable (Wang et al., 2016), there are many other lingering sources of structural uncertainty that could still be incorporated. For instance, generalized linear models could be used instead, which take a number of mathematical structures depending on the choice of link and response functions (Clyde 2003). Similarly, various data transformations could be used to generate competing models of the wildfire activities. Alternative models could be constructed that condense the two model components into VLF occurrence probabilities only, so that the event space of each month is purely binary. Instead of biogeographical classification of regions, the Continental United States could be partitioned using administrative or other boundaries to generate VLF predictions relevant to specific

stakeholders. Hence, clearly a broad variety of other structural uncertainties still exist that could potentially influence predictions of future VLF frequency in the second half of the 21st century.

It is important to understand that the VLF probabilities do not inform us as to what will actually happen, but rather communicates the degree of uncertainty about future outcomes conditional on carbon emission scenarios. For this reason, some tolerance to deviations between observed and expected VLF frequencies should be considered, as should the fact that the predictions were based on modeled climate data as opposed to direct observations. Still, in many regions, the ensemble performance was relatively adequate, and the simulated distribution of fire counts covered the observations. Moreover, when deviations occurred, they tended to underestimate the future VLF counts. Hence, the overall claim that VLF counts will increase in the future under climate change is supported by the results of this study, as well as through the work of others (Stavros et al., 2014, Barbero et al., 2015). Stochastic uncertainty will be critical when explicitly linking changes to VLF occurrence to human activities and for assessing the future levels of VLF simultaneity (Tedim et al., 2018) and is a factor that would be well addressed using the methods described here but is beyond the scope of this paper. The inherent stochasticity of the PET construction process suggests that repeated applications of this methodology in the future may yield slight variations to the results presented here.

Interestingly, a standard factor analysis revealed that more than 86 percent of the variability in predicted probabilities could be attributed to variance amongst the PETs rather than variance amongst the climate models, and while the PETs are an inherently unstable choice of predictive model, this suggests that structural uncertainties should receive the attention of climate impact researchers in much the same way that the choice of climate model does. Further exploration of these structural uncertainties in climate impact analyses cannot be recommended enough in future

analyses, as they inform not only of future impacts, but the reliability of these predictions, which can influence decision-maker behaviors in a variety of ways (Weber and Johnson 2009).

2.5 CONCLUSIONS

While the key conclusion from this research was that fires that were historically considered very-large and rare are likely to become increasingly frequent in most regions of the Continental United States at the end of the 21st century, there are also a number of other complexities in future wildfire activity that may be of further relevance to researchers and decision-makers. For instance, although temperature-based metrics were often important for prediction, this analysis also found that the identification of important predictors could be highly uncertain across a number of factors, which should be ignored at one's peril. Moreover, even using the relatively simple probabilistic models I developed, rich details regarding future wildfire activities were constructed that reasonably matched observed fire frequencies and were dynamic in terms of intra-annual trends, fire frequency, simultaneous fire occurrence, and the readiness with which LFs become VLFs. Although overall increases are predicted, I also observed exceptions and regional variability. In the Northwestern United States, VLF frequencies were predicted to increase, with nearly two additional events per year, and increases close to one additional VLF per year were fairly commonly throughout much of the Continental United States as well. In rare instances, the potential for decreases in VLF activity was also reported.

The cumulative impact of these changes is anticipated to affect decision-makers in various ways and the techniques described here have a number of benefits for addressing their needs. For instance, the presented Bayesian model averaging techniques avoids many of the risks of traditional model selection techniques that are especially dangerous when predicting complex

phenomena such as wildfire. Moreover, this method simultaneously provides a natural method of calculating important event probabilities that are critical to informed decision-making. While uncertainty in climate models is well understood amongst climate impact researchers, these results highlight the hidden sources of structural uncertainty, and encourage the use of Bayesian model averaging to reconcile them into robust forecasts of future wildfire and other impacts resulting from climate change.

2.6 REFERENCES

Achtemeier, G.L. On the formation and persistence of superfog in woodland smoke. Meteorological Applications. 2009, 16, 215–225.

Allen, C.D., Macalady, A.K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D.D., Hogg, E.T., & Gonzalez, P. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest ecology and Management*. 259(4), 660-684.

Arguez, A., & Vose, R.S. (2011). The definition of the standard WMO climate normal: The key to deriving alternative climate normals. Bulletin of the American Meteorological Society, 92(6), 699-704.

Arpaci, A., Eastaugh, C.S., & Vacik, H. (2013). Selecting the best performing fire weather indices for Austrian ecoregions. *Theoretical and Applied Climatology*, 114(3-4), 393-406.

Bailey, R.G. (2016). Bailey's ecoregions and subregions of the United States, Puerto Rico, and the US Virgin Islands. Forest Service Research Data Archive: Fort Collins, CO, USA, 2016. Available online: **https://doi.org/10.2737/RDS-2016-0003** (accessed on 14 December 2018).

Barbero, R., Abatzoglou, J.T., Kolden, C.A., Hegewisch, K.C., Larkin, N.K., & Podschwit, H. (2015). Multi□scalar influence of weather and climate on very large□fires in the Eastern United States. *International Journal of Climatology*. 35(8), 2180-2186.

Barbero, R., Abatzoglou, J.T., Larkin, N.K., Kolden, C.A., & Stocks, B. (2015). Climate change presents increased potential for very large fires in the contiguous United States. *International Journal of Wildland Fire*. 24(7), 892-899.

Barbero, R., Abatzoglou, J.T., Steel, E.A., & Larkin, N.K. (2014). Modeling very large-fire occurrences over the continental United States from weather and climate forcing. *Environmental Research Letters*. 9(12), 124009.

Barrett, K. The Full Community Costs of Wildfire. Headwaters Economics. Available online: https://headwaterseconomics.org/wp-content/uploads/full-wildfire-costs-report.pdf (accessed on 14 December 2018).

Bentz, B.J., Régnière, J., Fettig, C.J., Hansen, E.M., Hayes, J.L., Hicke, J.A., Kelsey, R.G., Negrón, J.F., & Seybold, S. J. (2010). Climate change and bark beetles of the western United States and Canada: direct and indirect effects. *BioScience*. 60(8), 602-613.

Beverly, J.L., & Bothwell, P. Wildfire evacuations in Canada 1980–2007. (2011) *Natural Hazards.* 59(1), 571–596.

Beverly, J.L., Flannigan, M.D., Stocks, B.J., & Bothwell, P. The association between Northern Hemisphere climate patterns and interannual variability in Canadian wildfire activity. *Canadian Journal of Forest Research.* 2011, 41, 2193–2201.

Bradley, B. A., Curtis, C. A., & Chambers, J. C. (2016). Bromus response to climate and projected changes with climate change. In Exotic Brome-Grasses in Arid and Semiarid Ecosystems of the Western US (pp. 257-274). Springer International Publishing.

Brooks, S.P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics.* 7(4), 434-455.

Brotak, E.A., & Reifsnyder, W.E. (1977). An investigation of the synoptic situations associated with major wildland fires. *Journal of Applied Meteorology.* 16(9), 867-870.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M. , Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*. 76(1).

Chen, J., Brissette, F.P., Poulin, A., & Leconte, R. (2011). Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. *Water Resources Research.* 47(12).

Climatology Lab. Available online: **http://www.climatologylab.org** (accessed on 29 January 2018).

Clyde, M. (2003). Model averaging. *Subjective and Objective Bayesian statistics*. 25, 320–326.

Crawford, J.A., Wahren, C.H., Kyle, S., & Moir, W.H. Responses of exotic plant species to fires in Pinus ponderosa forests in northern Arizona. *Journal of Vegetation Science.* 2001, 12, 261–268.

Dale, L. (2009) The True Cost of Wildfire in The Western US; Western Forestry Leadership Coalition: Denver, CO, USA.

Dennison, P.E., Brewer, S.C., Arnold, J.D., & Moritz, M.A. (2014). Large wildfire trends in the western United States, 1984–2011. *Geophysical Research Letters.* 41(8), 2928-2933.

Flannigan, M.D., Wotton, B.M., Marshall, G.A., De Groot, W.J., Johnston, J., Jurko, N., & Cantin, A. S. (2016). Fuel moisture sensitivity to temperature and precipitation: climate change implications. *Climatic Change.* 134(1-2), 59-71.

Development Core Team R. A Language and Environment for Statistical Computing. 2008. Available online: **http://www.R-project.org** (accessed on 14 December 2018).

Flannigan, M.D., Amiro, B.D., Logan, K.A., Stocks, B.J., & Wotton, B.M. (2006). Forest fires and climate change in the 21st century. *Mitigation and Adaptation Strategies for Global Change.* 11(4), 847-859.

Flannigan, M.D., Krawchuk, M.A., de Groot, W.J., Wotton, B.M., & Gowman, L.M. (2009). Implications of changing climate for global wildland fire. *International Journal of Wildland Fire.* 18(5), 483-507.

Forster, C., Wandinger, U., Wotawa, G., James, P., Mattis, I., Althausen, D., Simmonds, P., O'Doherty, S., Jennings, S.G., Kleefeld, C., & Schneider, J. (2001) Transport of boreal forest fire emissions from Canada to Europe. *Journal of Geophysical Research: Atmospheres.* 106(D19), 22887–22906.

Fragoso, T.M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review.* 86(1), 1-28.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of Markov Chain Monte Carlo.* 6, 163-174.

González-Cabán, A. Economic Cost of Initial Attack and Large-Fire Suppression. (1983). USDA Forest Service General Technical Report PSW-068; U.S. Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station: Berkeley, CA, USA, 7p.

Haffey, C., Sisk, T.D., Allen, C.D., Thode, A.E., & Margolis, E.Q. Limits to Ponderosa Pine Regeneration following Large High-Severity Forest Fires in the United States Southwest. *Fire Ecology.* 2018, 14, 143–163.

Hido, S., Kashima, H., & Takahashi, Y. (2009). Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal.* 2(5□6), 412-426.

Holsinger, L., Parks, S.A., & Miller, C. (2016). Weather, fuels, and topography impede wildland fire spread in western US landscapes. *Forest Ecology and Management.* 380, 59-69.

Krueger, E.S., Ochsner, T.E., Engle, D.M., Carlson, J.D., Twidwell, D., & Fuhlendorf, S.D. (2015). Soil moisture affects growing□season wildfire size in the Southern Great Plains. *Soil Science Society of America Journal.* 79(6), 1567-1576.

Littell, J.S., McKenzie, D., Kerns, B.K., Cushman, S., & Shaw, C.G. (2011). Managing uncertainty in climate□driven ecological models to inform adaptation to climate change. *Ecosphere.* 2(9), 1-19.

Mallick, B.K., & Gelfand, A.E. (1994). Generalized linear models with unknown link functions. *Biometrika.* 81(2), 237-245.

Marlon, J.R., Bartlein, P.J., Gavin, D.G., Long, C.J., Anderson, R.S., Briles, C.E. , Brown, K.J., Colombaroli, D., Hallett, D.J., Power, M.J., & Scharf, E.A. (2012). Long-term perspective on wildfires in the western USA. *Proceedings of the National Academy of Sciences.* 109(9), E535-E543.

Meyn, A., White, P.S., Buhk, C., & Jentsch, A. (2007). Environmental drivers of large, infrequent wildfires: The emerging conceptual model. *Progress in Physical Geography.* 31(3), 287–312.

Moeltner, K., Kim, M.K., Zhu, E., & Yang, W. (2013) Wildfire smoke and health impacts: A closer look at fire attributes and their marginal effects. *Journal of Environmental Economics and Management.* 66, 476–496.

Monitoring Trends in Burn Severity. Available online: **http://www.mtbs.gov** (accessed on 21 November 2017).

Monnahan, C.C., Thorson, J.T., & Branch, T.A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution.* 8(3), 339-348.

Morgan, M.G., Henrion, M., & Small, M. (1992). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press.

Neary, D.G., Gottfried, G.J., & Ffolliott, P.F. Post-wildfire watershed flood responses. In Proceedings of the 2nd International Fire Ecology Conference, American Meteorological Society, Orlando, FL, USA, 28 November–2 December 2003; Volume 65982.

Peppin, D.L., Fulé, P.Z., Sieg, C.H., & Beyers, J.L.; Hunter, M.E.; Robichaud, P.R. Recent trends in post-wildfire seeding in western US forests: Costs and seed mixes. (2011) *International Journal of Wildland Fire.* 20(5), 702–708.

Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 20–22 March 2003.

Podschwit, H., & Cullen, A. (2020). Patterns and trends in simultaneous wildfire activity in the United States from 1984 to 2015. *International Journal of Wildland Fire.* 29(12), 1057-1071.

Provost, F., & Domingos, P. (2000). Well-trained PETs: Improving probability estimation trees. Raport instytutowy IS-00-04, Stern School of Business, New York University.

Quinlan, J.R. (2014). *C4. 5: programs for machine learning*. Elsevier: Amsterdam, The Netherlands.

Raftery, A.E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association.* 98(464), 931-938.

Reid, C.E., Brauer, M., Johnston, F.H., Jerrett; M, Balmes, J.R., & Elliott, C.T. (2016) Critical review of health impacts of wildfire smoke exposure. *Environmental Health Perspectives.* 124, 1334.

Rocca, M.E., Miniat, C.F., & Mitchell, R.J. (2014) Introduction to the regional assessments: Climate change, wildfire, and forest ecosystem services in the USA. *Forest Ecology and Management.* 327, 8.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society.* 71(2), 319-392.

Stavros, E.N., Abatzoglou, J., Larkin, N.K., McKenzie, D., & Steel, E.A. (2014). Climate and very large wildland fires in the contiguous western USA. *International Journal of Wildland Fire.* 23(7), 899-914.

Stavros, E.N., Abatzoglou, J.T., McKenzie, D., & Larkin, N.K. (2014). Regional projections of the likelihood of very large wildland fires under a changing climate in the contiguous Western United States. *Climatic Change*, 126(3-4), 455-468.

Stephens, S.L., Burrows, N., Buyantuyev, A., Gray, R.W., Keane, R.E., Kubian, R., Liu, S., Seijo, F., Shu, L., Tolhurst, K.G., & Van Wagtendonk, J.W. (2014). Temperate and boreal forest mega□fires: characteristics and challenges. *Frontiers in Ecology and the Environment.* 12(2), 115-122.

Sitch, S., Huntingford, C., Gedney, N., Levy, P.E., Lomas, M., Piao, S. L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., & Jones, C.D. (2008). Evaluation of the terrestrial carbon cycle, future plant geography and climate□carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs). *Global Change Biology.* 14(9), 2015-2039.

Slocum, M.G., Beckage, B., Platt, W.J., Orzell, S.L., & Taylor, W. (2010). Effect of climate on wildfire size: a cross-scale analysis. *Ecosystems.* 13(6), 828-840.

Syphard, A.D., Sheehan, T., Rustigian-Romsos, H., & Ferschweiler, K. (2018). Mapping future fire probability under climate change: Does vegetation matter?. *PloS one.* 13(8), e0201680.

Syphard, A.D., Radeloff, V.C., Keeley, J.E., Hawbaker, T.J., Clayton, M.K., Stewart, S.I., Hammer, R.B. (2007). Human influence on California fire regimes. *Ecological Applications.* 17(5), 1388-1402.

Syphard, A.D., Keeley, J.E., Pfaff, A.H., Ferschweiler, K. (2017). Human presence diminishes the importance of climate in driving fire activity across the United States. *Proceedings of the National Academy of Sciences.* 144(52), 13750-13755.

Taylor, K.E., Stouffer, R.J., & Meehl, G.A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society.* 93(4), 485-498.

Taylor, S.W., Woolford, D.G., Dean, C.B., & Martell, D.L. (2013). Wildfire prediction to inform fire management: statistical science challenges. Statistical Science, 28(4), 586-615.

Tedim, F., Leone, V., Amraoui, M., Bouillon, C., Coughlan, M.R., Delogu, G.M., Fernandes, P.M., Ferreira, C., McCaffrey, S., McGee, T.K., & Parente, J. (2018). Defining extreme wildfire events: difficulties, challenges, and impacts. *Fire.* 1(1), 9.

Val Martin, M., Heald, C.L., Ford, B., Prenni, A.J., & Wiedinmyer, C. (2013) A decadal satellite analysis of the origins and impacts of smoke in Colorado. *Atmospheric Chemistry and Physics.* 13(15), 7429–7439.

Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics.* 17(1), 60.

Weber, E.U., & Johnson, E.J. (2009). Decisions under uncertainty: Psychological, economic, and neuroeconomic explanations of risk preference. In *Neuroeconomics* (pp. 127-144). Academic Press.

Weiss, A., & Hays, C.J. (2005). Calculating daily mean air temperatures by different methods: implications from a non-linear algorithm. *Agricultural and Forest Meteorology.* 128(1-2), 57-65.

Westerling, A.L. (2016). Increasing western US forest wildfire activity: sensitivity to changes in the timing of spring. *Philosophical Transactions of the Royal Society.* 371(1696), 20150178.

Westerling, A.L., Bryant, B.P., Preisler, H.K., Holmes, T.P., Hidalgo, H.G., Das, T., & Shrestha, S.R. (2011). Climate change and growth scenarios for California wildfire. *Climatic Change.* 109(1), 445-463.

Westerling, A.L., & Swetnam, T.W. (2003). Interannual to decadal drought and wildfire in the western United States. EOS, *Transactions American Geophysical Union.* 84(49), 545-555.

Williams, J. (2013). Exploring the onset of high-impact mega-fires through a forest land management prism. *Forest Ecology and Management.* 294, 4-10.

Zargar, A., Sadiq, R., Naser, B., & Khan, F.I. (2011). A review of drought indices. Environmental Reviews, 19(NA), 333-349.

Chapter 3. UNCERTAINTY IN UTILITY FUNCTION CHOICE FOR FORECASTS OF FIRE SIZE

3.1 INTRODUCTION

A wildfire's size changes over time. The largest and most destructive wildfires that have ever burned all began as small fires. As wildfire size increases, its impact on humans and the environment often also increases. Actively burning fire destroys property and can cause injuries and fatalities through both smoke inhalation and direct exposure to fire (Stephens et al. 2014). Smoke production (Moeltner et al. 2013), insured losses (Murname 2006), and firefighting costs (Murname 2006, González-Cabán 1983) all tend to be greater in larger fires than in smaller ones. Consequences to plant (Gebert et al. 2007, Sessions et al. 2004) and animal (Burton 2005, Rupp et al. 2006) populations can also be greater in large fires. Because impacts of wildfire are often correlated with fire size, the ability to predict the final size while it is still burning could help to anticipate and mitigate these risks.

Predicting wildfire size requires a model, for which there are many possible candidates. Wildfire simulators (Finney et al. 2011), built on spread models in combination with detailed information of fuel and fire perimeter, provide spatially explicit estimates of the location of future burning areas (Papadopoulos & Pavlidou 2011). These models of wildfire growth, along with estimates of fire duration, provide estimated maps that could be used to infer final fire size (Finney et al. 2011, Papadopoulos & Pavlidou 2011) Although these complex models are preferred for informing firefighting decisions, many of the physical models can be computationally expensive (Taylor et al. 2013). Alternatively, wildfire size predictions might also come from much simpler statistical models informed by environmental controls on fire size. For example, machine-learning methods have been applied to predict final fire size in Boreal

Alaska using information available at the time of ignition (Coffield et al. 2019). Regression techniques have also been applied in ways that could infer final fire size. For instance, weeks in which very large wildfires are most likely to occur can be reasonably predicted in a generalized linear model framework (Stavros et al. 2014, Barbero et al. 2013).

In practice, users of fire size predictions must select some subset of the universe of potential model choices from which to produce forecasts. Selection of the subset is commonly done by identifying a single model that optimizes a specified performance measure. This method of subset selection is simple, but it fails to consider three issues critical to decision makers: (1) the specified performance measure may not represent the preferences of all model users; (2) data availability and quality can change over time, and the "best" forecast model might be useful only at a specific point in time; (3) the degree to which two or models agree is itself important information, and this information is not available from single models.

No one model can be best at all things, so resolving issue 1 requires multiple models. When using models of wildfire size, decision makers have varying objectives. For example, one decision maker might need predictions that are, on average, close to observed values, while another decision maker might need to correctly predict the outcome of extreme events as soon as they begin. A third decision maker might need to accurately reproduce historical data. Yet another decision maker may instead need a model to be good at multiple performance measures simultaneously and seek a Pareto optimal choice of model (Hummel et al. 2013). Multiple models, each optimized for unique performance measures, would represent the diversity of preferences about model quality better than any single model (Willmott 1982, Morgan et al. 1992).

With regard to issue 2, multiple-model approaches can also account for changes in the availability and quality of data over time. For example, it is common to report both a full (optimal) model and a reduced (less complex) model. The two options provide model users the choice to generate predictions from either the optimal model when the necessary data is available, or from the reduced models when certain quantities relevant to prediction may be unavailable, highly uncertain, or impractical to calculate (e.g. Thies et al. 2006). Additionally, if the relevance of factors used in prediction are time-varying, then multiple models may better optimize predictive performance. As an example in the context of wildfire, consider the time-varying relevance of soil moisture on fire size in the Great Plains, where the availability of soil water has one relationship with wildfire occurrence in the growing season (May-October) and another relationship with wildfire occurrence in the dormant season (November-April) (Krueger et al. 2015). Time-varying covariates are also seen at the scale of an individual wildfire. For example, in a study of wildfire size in south-central Florida, wind disproportionately influenced the growth of smaller fires, but drought intensity was a stronger influence on growth once the fire's exceeded certain size thresholds (Slocum et al. 2010).

The extent to which multiple appropriate models agree on an outcome (i.e. issue 3) is especially useful information when making costly decisions, as is often the case in real-time wildfire management. Predictions from multiple models provide a range of plausible scenarios of future quantities (e.g. climate models; Littell et al. 201), which can be combined into a single prediction or distribution with a metamodeling approach, such as Bayesian Model Averaging (Hoeting et al. 1999). Whichever approach is adopted, if the choice of model radically changes predictions, then decision makers should have less confidence in its overall assessment than if all the models were in general agreement. When and where model coherence is low, decision

makers would be encouraged to adopt risk-averse choices, particularly when the impacts are potentially severe. Hence, multiple-model approaches can provide one type of safeguard against otherwise overconfident decisions.

In this study, I explored these issues in the context of predicting the size of individual wildfire size. Others have attempted this task in the past, namely (Coffield et al. 2019) who used decision trees to predict fire size on the ignition data in Boreal Alaska. However, our models will also explore at least four aspects that were not examined in (Coffield et al. 2019). First, our models here are applied to a suite of new geographic regions, not just Alaska. Second, our models will use wildfire growth information available later in the wildfire's lifetime to attempt to improve the accuracy of the predictions. Third, the models presented here optimize multiple performance measures. Finally, models presented here are built on simple regression techniques rather than machine learning, which have the advantage of being easy to interpret and relatively simple for non-statisticians to apply. I use the models I developed to answer the following questions:

A. How might preferences of wildfire managers regarding the choice of model performance measure affect predictions of fire size?
B. Can predictions of individual fire size be improved as more information becomes available?
C. How well do statistical models of fire size compare to existing wildfire spread models?

I developed two classes of fire size models that predicted fire size during two unique stages of a wildfire's lifetime: first-day model classes and accumulated evidence classes. The first-day class of models used information available on an incident's ignition day to predict fire size. The accumulated evidence class of models used information about the wildfire's growth, which is

available later on in the wildfire's lifetime, to produce refined predictions. To reflect the potentially diverse preferences of wildfire managers of what is considered a good model, the models were selected according to three performance measures that each represent a particular decision-making need. This set of models I refer to as the triple criteria model set (TCMS). For fire-prone regions in the Continental United States (CONUS), I identified the TCMS for each of the two model classes and, therefore, produced up to six unique models. I compared the predictions and structure of models across each TCMS and found that the optimal model structure depended on the location and choice of performance measure. The predictive performance of first-day models was sometimes improved by using wildfire growth information, and the simple statistical models contained in the TCMS was sometimes competitive with complex wildfire spread models currently used in firefighting.

3.2 METHODS

3.2.1 *Burned area data*

I used burn area data from three sources: Monitoring Trends in Burn Severity (MTBS), ICS-209 reports, and InciWeb website reports. Each data source met unique objectives of this research that could not be accomplished with the others.

Data from MTBS (Eidenshink et al. 2007) provide estimates of individual fire size, which I used to fit the first-day class of models. The dataset included events within the CONUS and for the years 1984-2016. Ignition day was assumed to be the discovery date.

ICS-209 reports (https://fam.nwcg.gov/fam-web/hist_209/report_list_209, accessed August 2015) and InciWeb (https://inciweb.nwcg.gov/) provide estimates of daily fire size, which I used to fit and validate the accumulated evidence class of models. The ICS-209 data span all of CONUS for the years 2002-2013. InciWeb website data provide daily fire size

information from June 2 to December 31, 2018. The reported time series was reconstructed and cross-referenced with Incident Management Situation Reports (https://www.predictiveservices.nifc.gov/intelligence/archive.htm) and other data sources to ensure that the growth time series were as accurate as possible. InciWeb growth time series data were used to independently validate both the first-day and accumulated evidence classes of models. For both ICS-209 and InciWeb data, a burn area time series the ignition date was either inferred from the reported size of the time series or explicitly provided.

All three data sources were prescreened to remove events that were either 1000 acres or smaller, unnamed, or not classified as a wildfire. To maintain consistency between datasets, InciWeb and ICS-209 data were also screened to remove wildfire complexes. Each wildfire was assigned to one of 19 Bailey's biogeographical regions (Figure 3.1). Of these, the Marine, Hot Continental Mountains, Subtropical Mountains, and Warm Continental Mountains regions were dropped from analysis because they lacked large fire events, leaving 15 biogeographical regions in the analysis (Table 3.1).

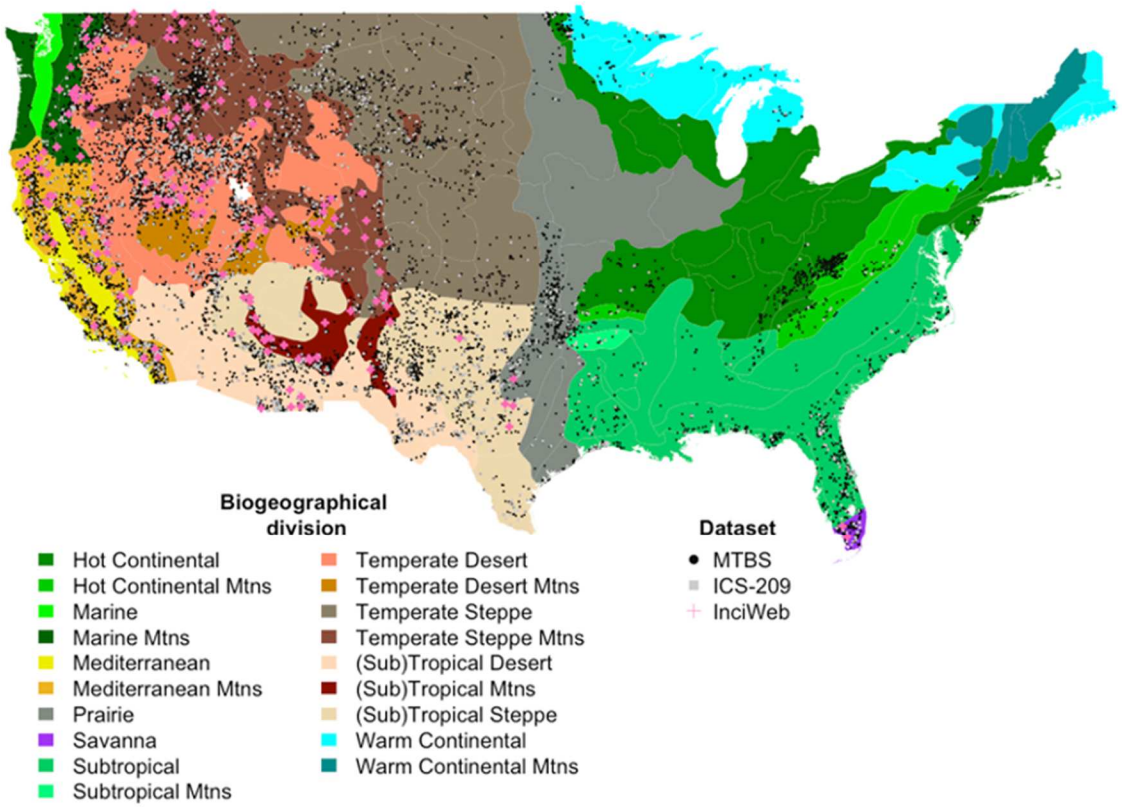


Figure 3.1. Map of Bailey's ecoregions at division level overlaid with fire location data from MTBS, ICS-209, and InciWeb. Of the 19 total regions in the continental United States, 15 had sufficient data to be used in the study. Excluded regions: Marine, Hot Continental Mountains, Subtropical Mountains, and Warm Continental Mountains regions.

Table 3.1. Number of wildfires for which data were available across each of three datasets, Monitoring Trends in Burn Severity (MTBS), Incident Command System 209 reports (ICS-209), and InciWeb website updates, for each of the 15 ecoregions analyzed.

Biogeographical region	MTBS (1984-2016)	ICS-209 (2002-2013)	InciWeb (2018*)
Hot Continental	362	14	0
(Sub)Tropical Steppe	969	133	9
Temperate Desert	2523	332	36
Temperate Steppe Mtns	1388	264	47
Temperate Desert Mtns	213	49	9
(Sub)Tropical Desert	640	103	5
Mediterranean Mtns	1208	197	29
Mediterranean	220	30	1
Temperate Steppe	850	89	0
Marine Forest Mtns	272	46	15
Subtropical	769	73	0
Prairie	468	42	1
(Sub)Tropical Mtns	344	58	17
Warm Continental	95	10	0
Savanna	136	19	2
* From June 2 to December 31, 2018.			

3.2.2 *Covariates*

Wildfire size was predicted using a combination of meteorological, environmental, and fire growth data. Meteorological information was collected from gridded weather data from the

University of Idaho gridMET project (http://www.climatologylab.org/gridmet.html). I considered nine meteorological covariates that could plausibly be related to wildfire size: 100-hour fuel moisture, 1000-hour fuel moisture, the model-G energy release component, burning index, daily average temperature, specific humidity, potential evapotranspiration, vapor pressure deficit, and wind speed. Daily average temperature was calculated as one-half the sum of the daily maximum and minimum values (Weiss and Hays 2005).

In addition to these meteorological covariates, which vary over time, three time-invariant environmental covariates were mapped to each individual fire: elevation, topographic roughness (Riley et al. 1999), and the human influence index (Sanderson et al. 2002). The elevation measures were calculated using elevation data from the gridMET project (http://www.climatologylab.org/gridmet.html). Topographic roughness was calculated using a neighborhood of all eight adjacent pixels. The human influence index measures the level of human activity in the vicinity of the fire and was included to distinguish among wildfire activity in urban, wildland interface, and wilderness areas.

The accumulated evidence models incorporated wildfire growth covariates calculated from the daily wildfire size time series. Specifically, the wildfire growth covariates measured the presence/absence of a large growth event, where a large growth event was defined as a daily growth increment that exceeded one of three absolute thresholds (k) of fire size: 1000 acres, 2500 acres, or 10000 acres.

Each covariate was classified into one of seven dimensions of the fire environment: fuel moisture content, atmospheric moisture content, temperature, wind, topography, anthropogenic activity, and wildfire growth (Table 3.2). First-day models used covariates from six dimensions instead of seven because, by definition, wildfire growth covariates were not included.

Table 3.2. The 15 covariates considered in model selection and the associated seven dimensions of the fire environment.

Covariate	Dimension
100-hour fuel moisture	Fuel moisture content
1000-hour fuel moisture	Fuel moisture content
Energy release component	Fuel moisture content
Burning index	Fuel moisture content
Potential evapotranspiration	Fuel moisture content
Specific humidity	Atmospheric moisture content
Vapor pressure deficit	Atmospheric moisture content
Average temperature	Temperature
Wind speed	Wind
Elevation	Topography
Topographic roughness	Topography
Human influence	Anthropogenic factor
1000 acres of growth	Wildfire growth
2500 acres of growth	Wildfire growth
10000 acres of growth	Wildfire growth

3.2.3 *Model candidates*

Predictions came from generalized linear models, which are defined by a link function and probability distribution. The link function transforms a linear combination of covariates into the response, fire size, and the probability distribution describes the error structure of the model. I assumed that the errors were Gamma-distributed and considered two possible links, logarithmic and inverse functions.

Although the first-day and accumulated evidence models can both produce estimates of final size of fire ($\widehat{K}$), they use separate methods to calculate this quantity. The first-day models estimate $\widehat{K}$ directly, so that,

$$\widehat{K} = 1000 \times \exp(f(X)). \qquad (3.1)$$

Here, X is a vector of relevant environmental information, and f is the generalized linear model functional form.

The accumulated evidence models produced estimates of $\widehat{K}$ indirectly by instead estimating the difference between the current size and final size. The difference was calculated with either additive or multiplicative methods. The additive method estimated $\widehat{K}$ by predicting the remaining amount of area that will after the forecast day.

$$\widehat{K} = \exp(f(X)) + A(t^*). \qquad (3.2)$$

Here, $A(t^*)$, represents the size of the fire on the forecast day, t^*. The multiplicative method instead estimated the factor between the size on the forecast day, $A(t^*)$, and final size.

$$\widehat{K} = \exp(f(X)) \cdot A(t^*). \qquad (3.3)$$

These model forms imply two responses for the accumulated evidence models.

$$y_+ = \ln(K - A(t_+^*)). \quad (3.4)$$

$$y_\times = \ln(K/A(t_\times^*)). \quad (3.5)$$

The accumulated evidence models make either one or two forecasts of final fire size. The first, mandatory, forecast is produced on the first day. This forecast assumes that no large wildfire growth events (i.e. a daily growth increment that exceeds one of the three specified thresholds) have occurred and, like the first-day model, uses environmental information associated with the

ignition day to produce an initial prediction of fire size. However, if a large growth event occurs, a second forecast is produced that uses the environmental information on the day of the large growth event. The additive models define current size using measurements at the beginning of each day, and the multiplicative models define current size using measurements at the end of each day. This distinction allows the multiplicative model to be used on the first day of a fire, $(t = 1)$, when the current size is zero, $A(t)$=0. Hence, the relevant forecast day, t^*, was defined as

$$t_+^* = \max(\{2, \min(\arg\max 1_{\{A(t+1)-A(t)>k\}})\}); \qquad (3.6)$$

$$t_\times^* = \max(\{1, \min(\arg\max 1_{\{A(t)-A(t-1)>k\}})\}). \qquad (3.7)$$

In some cases, the size associated with the second forecast day was missing and a significant growth event was known to have occurred over some interval, but the specific day of the growth event was unknown. In this case, the next day in which a burn area measurement was available was used as the forecast day instead. All three model classes are summarized in Table 3.3.

Table 3.3. Summary of the general structure and data sources used for each model class. Burn area data were used to calculate the response variables, and the covariates groups were used to generate forecasts of response. Wildfire growth is included in the accumulated evidence model class, but not the first-day model class.

Model class	Response variable	Burn area data	Covariate groups
First-day	Total burn area minus 1000 (log-transformed)	Monitoring Trends in Burn Severity (MTBS)	Atmospheric moisture Fuel moisture content Temperature Wind speed Topography Anthropogenic factors

Accumulated evidence (additive growth)	Burn area minus size of first large growth event or zero (log-transformed)	ICS-209	Atmospheric moisture Fuel moisture content Temperature Wind speed Topography Anthropogenic factors Wildfire growth
Accumulated evidence (multiplicative growth)	Burn area divided by size at time of large growth event or one (log-transformed)	ICS-209	Atmospheric moisture Fuel moisture content Temperature Wind speed Topography Anthropogenic factors Wildfire growth

3.2.4 *Triple Criteria Model Set (TCMS)*

To represent diverse preferences of wildfire managers, I selected models according to three performance measures, which generated a set of models for each class that I refer to as the triple criteria model set (TCMS). Each one of the models in the TCMS was selected from a large initial suite of candidate models to optimize one of three performance measures. The structure of all models can be described with nine model components: the link function, the method, and one model component for each of the seven potential dimensions. The initial model suite considered all possible configurations of these nine model components. The initial suite for the first-day class of models considered 864 models; i.e., all 432 linear combinations of all six dimensions, using both the logarithmic and inverse link functions. The initial suite for the accumulated

evidence class of models contained 5184 models; i.e., all 1296 linear combinations of all seven dimensions (including the mandatory wildfire growth dimension), using both the log and inverse functions, and using the additive and multiplicative methods.

Predictive performance was assessed with a combination of three measures of model quality: mean multiplicative error, recall of large fires, and likelihood. Mean multiplicative error is the mean of the absolute value of log-difference between the observations and predictions. Recall is the mean predicted probability of a fire exceeding 5000 acres when the fire eventually did exceed 5000 acres. Likelihood is the probability that the testing data set was produced from the generalized linear model.

The predictive performance of each set of generalized linear models was estimated from an initial suite of candidates using a Monte Carlo cross validation (MCCV) approach (Piccard and Cook 1984) that preserved the approximate size distribution for every iteration of the simulation. The MCCV simulation was performed as follows. First, each wildfire incident was assigned to one of four acreage classes: 1000- 2500, 2501-5000, 5001-15000, or >15000 acres. At each iteration of the cross-validation procedure, an equal number of random samples were drawn, without replacement, from each acreage class to produce training and testing datasets. If an acreage class contained an odd number of events, n, then the testing data set had one more observation than the training dataset. The training dataset was then used to fit each model in the initial model suite, and the testing dataset was used to assess the model's predictive performance.

Convergence of the MCCV simulations was determined by testing for a statistically detectable difference between the predicted performance of the first- and second-best models (Table 3.5). Specifically, a two-sampled t-test was used to test the null hypothesis that the differences between the performance measures of the best model and its closest competitor are

different from zero. In addition to the MCCV procedure (Table 3.6), I also estimated predictive performance using holdout validation of the InciWeb dataset (Table 3.7), which provides more realistic estimates of predictive performance than MCCV.

The three best models in the TCMS were selected by choosing the elements of the initial model suite that optimized the three performance measures (Table 3.8, Table 3.9). I evaluated which model components (covariates, link function, method) were always included in the TCMS, regardless of the choice of performance measures, which will I hereafter refer to as consensus components.

The level of model uncertainty was represented using the model range. The model range of a fire was defined as the ratio of the largest and smallest predictions of fire size across the three possible predictions of the TCMS. This quantity represents the maximum change in fire size predictions that could result from a model user changing preferences. In addition to using MCCV to assess model performance, in some divisions, I was able to compare the expected error estimates of the statistical models to expected error estimates of FSPro, a wildfire simulator. That is, the distribution of error levels of between expected sizes and actual sizes from FSPro was compared to those calculated from the error-optimizing model of the TCMS. All statistical analyses were performed using the R programming language (R Core Team 2018).

3.3 RESULTS

3.3.1 *Sensitivity to uncertainties in preferences*

It was rare that the same set of covariates were selected in the three models of the TCMS, each of which optimized a unique performance measures (errors, recall, likelihood). For example, in the Savanna region using the first-day models, the fuel moisture content dimension

was measured three ways: (1) 1000-hour fuel moisture in the model optimizing expected errors; (2) 100-hour fuel moisture in the model optimizing recall; (3) omitted in the model optimizing the out-of-sample likelihood (Figure 2). The choice of link function, as well as the method used to estimate the final fire size in the accumulated evidence model class, also changed depending on the choice of performance measure (Table 3.8, Table 3.9, Figure 3.2).

The accumulated evidence model had seven dimensions available compared to six in the first-day models, but still tended to use fewer dimensions than the first-day models (Figure 3.2). In only five cases did the accumulated evidence model use more covariates than the corresponding first-day models (Figure 3.2).

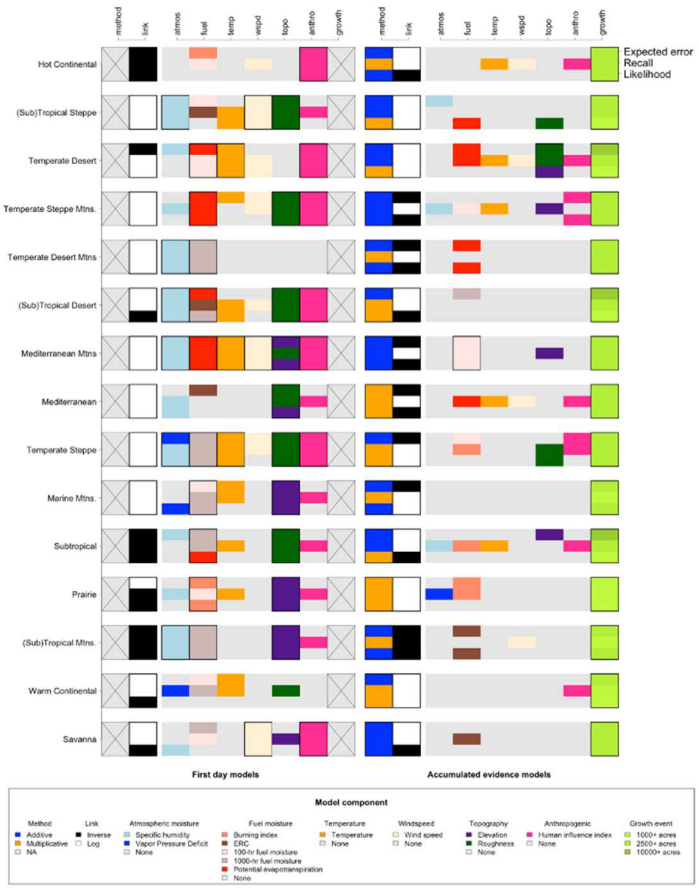


Figure 3.2. Summary of which model covariates were used in the elements of the triple criteria model set for each biogeographical region. X-axis (top): the nine model components in each of two model classes. Y-axis: the 15 biogeographical regions. For each region, three parallel horizontal bars show each of the three best models optimizing the performance measures of error, recall, and likelihood. Model dimensions common to all of three best models are identified with a black outline (e.g., fuel moisture content dimension in (Sub)Tropical Desert).

Across each TCMS, consistent use of dimensions was far more common than consistent use of particular covariates. For example, in the first-day models, regardless of the performance measure optimized, fuel moisture content was commonly included in the models (Figure 3.3), but the specific choice of covariate was highly variable (Figure 3.2). The number of models within each TCMS using a particular dimension, was usually lower in the accumulated evidence models compared to the first-day models (Figure 3.3).

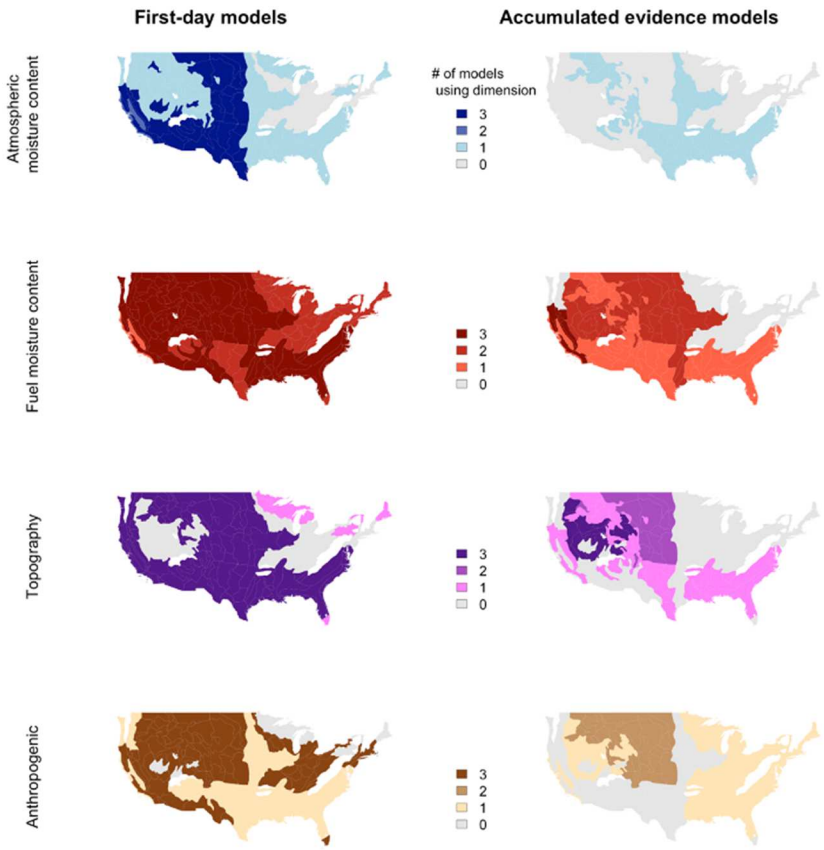


Figure 3.3. Map of the number of models in the triple criteria model set (0, 1, 2, 3) that used a particular dimension (atmospheric moisture content, fuel moisture content, topography, and anthropogenic)). The accumulated evidence model tended to use these dimensions less frequently than in the first-day models. Only biogeographical regions with multiple covariate choices within a dimension are shown. Temperature and wind speed dimensions were omitted because they each only contained one possible covariate. The wildfire growth covariates were omitted because they were relevant only to the accumulated evidence model class.

Models within the first-day model class tended to closely agree; the model range rarely exceeded a factor of two. For the first-day class, the minimum median model range occurred in the Temperate Desert Mountain region (1.02) and the maximum median model occurred in the Savanna region (1.33). Within the accumulated evidence model class, the predictions were more variable; the model range commonly exceeded a factor of two. For the accumulated-evidence class, the minimum median model range occurred in the Mediterranean Mountains region (1.04) and the maximum median model occurred in the Warm Continental region (3.21). In some cases,

in both the first-day and accumulated evidence classes, this model range was reported on the scale of hundreds. In the Mediterranean region, the predictions from models within the first-day class TCMS differed by a factor of 272 in the worst-case scenario. In the (Sub)Tropical Mountains, the predictions from the models within the accumulated-evidence class TCMS differed by a factor of 269 in the worst-case scenario (Figure 3.4).

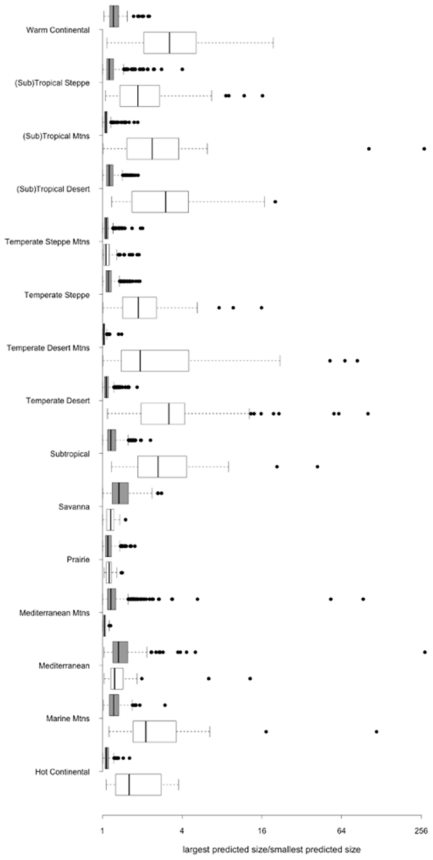


Figure 3.4. The model range (the largest prediction divided by the smallest prediction) of each fire event for each of the 15 biogeographical regions, as assessed with Monte Carlo cross validation. The ranges of predictions are shown for the first-day models (gray) and the accumulated evidence models (white).

3.3.2 *Predictive performance*

As estimated via MCCV, the accumulated evidence models always had smaller average multiplicative errors than did the first-day models. Although the accumulated evidence model

was favored under both validation techniques, larger estimated errors occurred when calculated with the holdout method: the median error ratio across biogeographical regions was a factor of 0.64 under MCCV and 0.88 under holdout (Figure 3.4). Under the holdout method, the (Sub)Tropical Steppe, Savanna, Marine Mountains regions had first-day models with smaller average errors than the accumulated evidence model classes (Figure 3.5).

Average recall scores (the predicted probability that a large fire will occur when a large fire eventually does occur) were always larger in the accumulated evidence model classes than in the first-day model classes, regardless of the biogeographical region or choice of validation method. The predicted probabilities were generally two to three times higher in the accumulated evidence model than in the first-day models, with only slight differences between the estimates obtained from MCCV versus holdout: median increase in recall across regions was a factor of 2.67 under MCCV and 2.76 under holdout (Figure 3.5). In no case was the recall larger in the first-day models than in the accumulated evidence models.

The out-of-sample log-likelihood was usually higher in the accumulated evidence models compared to the first-day models. Under MCCV, the optimal accumulated evidence models outperformed the optimal first-day model class in 13 of the 15 regions. Only in the Subtropical and Warm Continental division did the optimal first-day model better describe the distribution of fire size than the optimal accumulated evidence model. Similar results were observed under the holdout method. In eight of the 11 regions, the optimal accumulated evidence model outperformed the optimal first day model. Only in the (Sub)Tropical Mountains, (Sub) Tropical Steppe, and Mediterranean regions did the first day model outperform the accumulated evidence models (Figure 3.5). The absolute predictive ability of the TCMS models is in the Appendix (Table 3.6, Table 3.7).

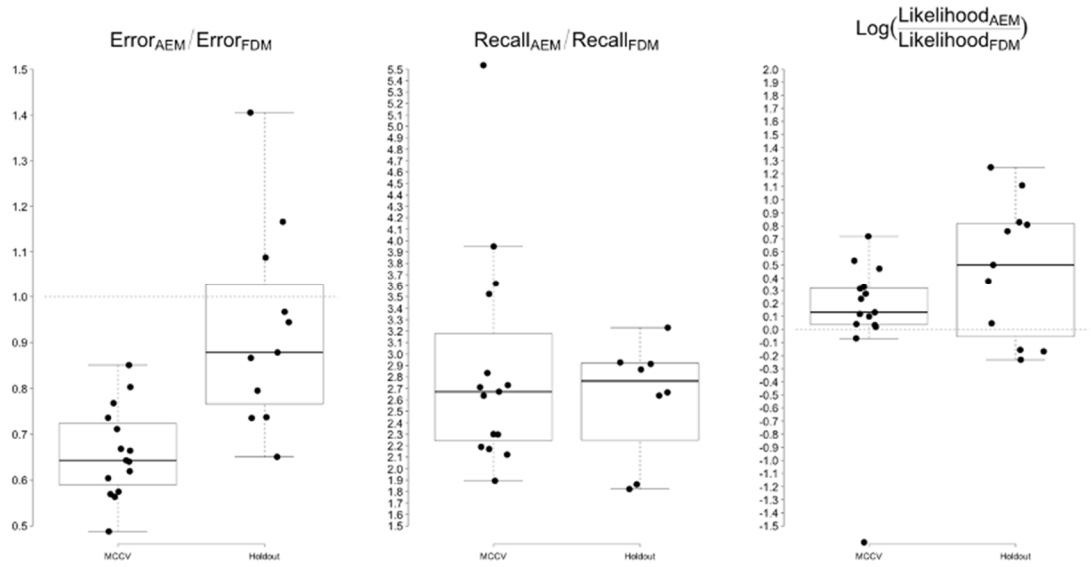


Figure 3.5. Differences in mean predictive performance (expected errors, recall, and likelihood) between accumulated evidence models (AEM) and first-day models (FDM). Estimates of predictive performance were obtained via Monte Carlo cross validation (MCCV) and holdout. When expected errors and recall were used as performance measures, AEMs had more favorable predictive performance than FDMs. On the other hand, when likelihood was used as a performance measure, FDMs had more favorable predictive performance than AEMs. For many regions, then, either model class could be preferred depending on the performance measure selected.

3.3.3 *Expected errors: TCMS versus FSPro*

Although the number of FSPRo runs were small for some regions, the cumulative distribution plots of the multiplicative factors separating observations and predictions of fire size suggested that the TCMS models could be competitive with FSPro (Figure 3.6). Most notably, in the Temperate Steppe Mountains, the expected errors of the TCMS models from both classes were lower than those obtained from FSPro. In other cases, FSPro outperformed the statistical models only slightly. For example, in the (Sub)Tropical Steppe, the expected errors in fire size predictions were nearly identical. In the Temperate Desert, FSPro had the lowest expected errors, but was closely matched by the performance in the accumulated evidence model. In the

remainder of regions, FSPro outperformed the statistical models, by a multiplicative factor in the range of 2-4. See Appendix (Table 3.10).

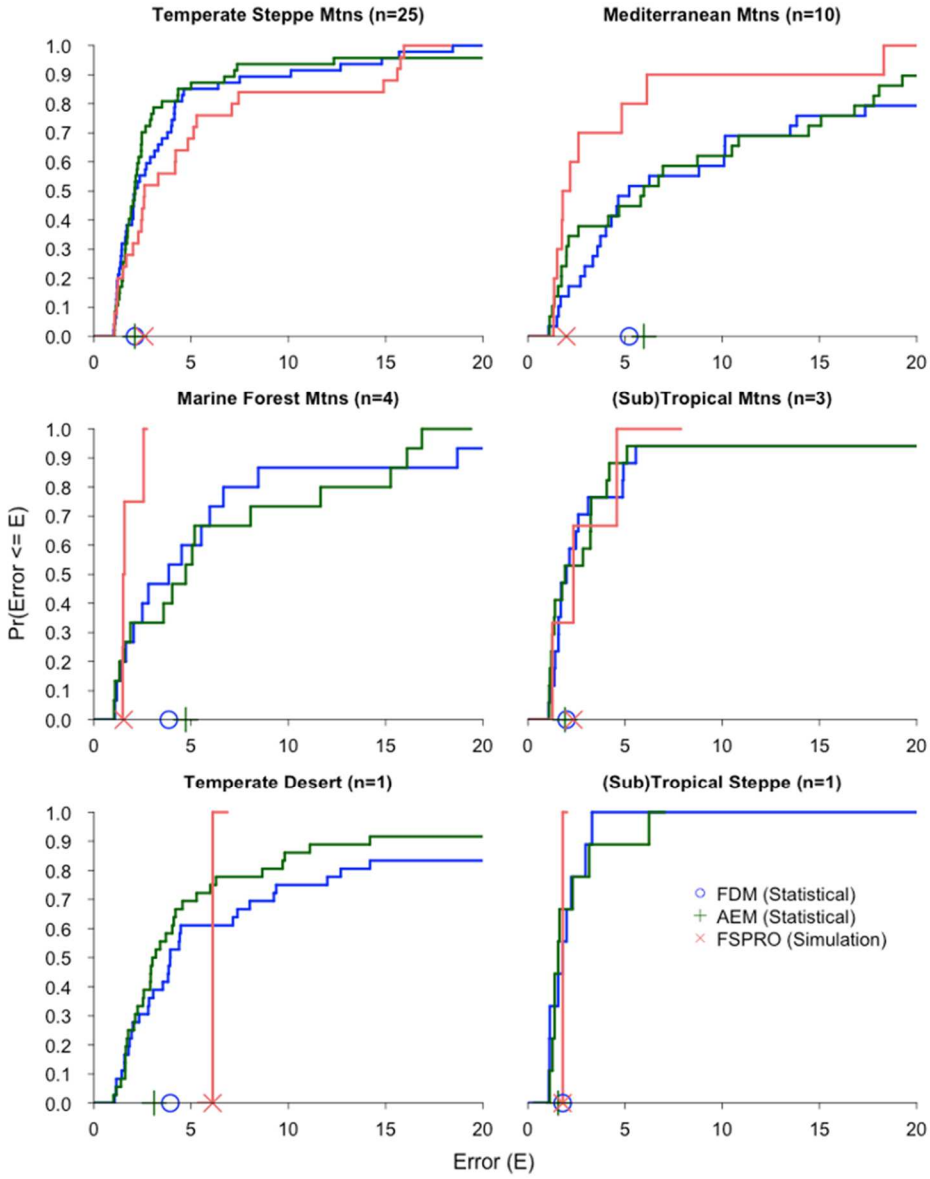


Figure 3.6. Empirical cumulative distribution function for errors in three classes of models: first-day models (FDM), accumulated evidence models (AEM), and Fire Spread Probability (FSPRO) in six geographical divisions. X-axis: median error levels of the three model classes. Y-axis: frequency of observations with error levels equal to or less than certain threshold. The curves identify the percentage of wildfires in the holdout data set that were predicted within a specified error tolerance. The median error levels of the three model classes are identified on the x-axis. The relative predictive performance of the statistical models compared to complex wildfire simulators varied by location, with wildfire simulators preferred in Mediterranean California and Marine Forest Mountains. The number of fires validated by FSPRO are identified in the title of each plot.

3.4 DISCUSSION

3.4.1 *Model uncertainty and user preference*

Observed relationships between environmental covariates and wildfire size will be sensitive to both selection of model performance measure and the time since ignition. A single covariate was unlikely to be included for all elements of the TCMS and also unlikely to be used in both the first-day and corresponding accumulated evidence model. However, the chances all elements of the TCMS contain a covariate drawn from a similar set, such as fuel moisture covariates, is higher. Hence, covariates identified as "best" are likely to change when new performance measures are considered, but the overall structure can be relatively stable to these changes. This means that although the specific structure of the best models is almost certain to change when new user needs are considered, the models will likely have common themes (e.g. always including drought indicators). This also means that selecting the best model for a particular application requires carefully identifying and defining how performance is measured. It is also worth noting that some measures are sensitive to specific kinds of behaviors. For instance, sensitivity is relevant to wildfire events in the extremes of the size distribution. In addition to uncertainties in model preferences, the time since ignition introduces additional uncertainty regarding the optimal choice of covariates. Fuel moisture content and topography frequently selected dimensions in the first-day model class: they were consensus dimensions in 10 of the 15 divisions. However, in the accumulated evidence model class, the fuel dimension was a consensus dimension in the Mediterranean Regime Mountains only, and topography was a consensus dimension in the Temperate Desert only (Table 3.4). Hence, as the wildfires grow, the relative importance of environmental covariates can decrease, and growth information can instead become more informative of final fire size.

The effects of uncertain model preferences can extend beyond uncertainties in the optimal model structure, as predictions from the models could also be highly variable. For both model classes, there existed cases where a reconsideration of preferences could change predictions by a factor of hundreds (Figure 3.4). Although these are typically extreme cases, the potential for this magnitude of change in predictions is relevant because the preferences of model users are themselves often highly uncertain, and the effects of selecting incorrectly is apparently highly consequential to the predictions of fire size. If these predictions are informing individuals making costly decisions, then this variability in predictions is highly important, as they could imply taking drastically different decisions.

The sensitivity of model structures and resulting predictions to uncertainties in model preferences has at least two implications. First, model users should not base decisions on the predictions of a single model unless they have good evidence that the model accurately reflects their preferences, which I suspect is seldom the case. Second, there are hidden risks associated with attempts to identify a single best model of fire size, as what is best can change across individuals and time. In our study, covariates were often added, dropped, or substituted when new performance measures were adopted (see Section 3.1). I thus recommend that researchers reframe questions of "Which models are best?" as "Are these predictions robust to the uncertainties in the preferences of model users?"

Table 3.4. Summary of consensus models for each of the 15 biogeographical divisions and two model classes (first-day and accumulated evidence). Only the fuel moisture content dimension was common to both model classes, and the final column identifies for which divisions this is true. The models used seven possible dimensions: anthropogenic (anthro), atmospheric moisture content (atmos), fuel moisture content (fuel), wind, temperate (temp), topographic (topo), and wildfire growth covariates (growth).

Biogeographical division	First-day model	Accumulated evidence model
Hot Continental	anthro	growth
(Sub)Tropical Steppe	atmos+wind+topo	growth
Temperate Desert	fuel+temp+anthro	topo+growth
Temperate Steppe Mtns	fuel+topo+anthro	growth
Temperate Desert Mtns	atmos+fuel	growth
(Sub)Tropical Desert	atmos+fuel+topo+anthro	growth
Mediterranean Mtns	fuel+temp+wind+topo+anthro	fuel+growth
Mediterranean	topo	growth
Temperate Steppe	atmos+fuel+temp+topo+anthro	growth
Marine Forest Mtns	fuel+topo	growth
Subtropical	fuel+topo	growth
Prairie	fuel+topo	growth
(Sub)Tropical Mtns	atmos+fuel+topo	growth
Warm Continental	none	growth
Savanna	wspd+anthro	growth

3.4.2 *Comparison of accumulated evidence and first-day models*

The predictive ability of initial forecasts of fire size could sometimes be improved by including data on the fire's growth. The performance measures of expected errors, recall of large fire, and out-of-sample likelihood were usually more favorable in the accumulated evidence models than in the first-day models. However, despite the often-favorable predictive ability of the accumulated evidence models, less accurate predictions from the first-day models may sometimes still be necessary because of data availability and quality considerations. For example, in the first days of a fire, whether a wildfire will have a large growth event exceeding a critical threshold is unknown. Similarly, if statistical models are applied to climate model data to project wildfire activity, the covariates needed to use the accumulated evidence model are by

definition unavailable. In both of these examples, the first-day model would be the better choice, but only because of data constraints, not superior predictive ability.

In general, the first-day models were more complex and used more dimensions than the accumulated evidence models, but had worse predictive performance compared to the accumulated evidence models. In other words, more variables were used to make worse predictions. This result suggests that much of the information in the "dropped" covariates in the first-day models could often be captured with wildfire growth covariates.

3.4.3 *Relative performance of TCMS and FSPro*

The predictive performance of the simple statistical models used in the TCMSs, in some contexts, was either competitive or outperformed the wildfire simulator FSPro (Table 3.10). FSPro uses detailed data inputs from many sources to generate expected fire size, including rasterized historical and future weather, fuel bed maps, and fire perimeter data. FSPro also requires that data inputs be selected by the model user, such as burn period duration and the probability of spot fire. In contrast, our most complex statistical models used only daily weather information, topographic, and human land use data at a particular location and day, as well as wildfire growth information, to predict expected fire size. These differences in model complexity result in large differences in run time. FSPro may require hours to complete, depending on user parameter choices, but the simple statistical models such as ours produced near-instantaneous predictions. Of course, FSPro does much that these statistical models do not, such as producing maps of fire extent. But if model quality is gauged by the difference between predictions and observations of fire size, our results suggest that the information requirements of FSPro could be reduced with little difference in relative performance. This comparison also suggests that statistical models such as ours can complement predictions from spatially explicit

models to verify that the forecasted fire spread is plausible, based on historic statistical relationships.

3.4.4 *Limitations*

Reducing uncertainty in predictions is desirable, because we would make better decisions if we knew the future with less error. However, when predictions come from a single model, uncertainty is artificially reduced, creating a false sense of certainty to model users which can encourage risk taking. The methods in this study accounted for multiple uncertainties when predicting wildfire size that are typically not evaluated in wildfire research and applications. Although I addressed some sources of uncertainty, a potential shortcoming of TCMS is that further uncertainties remain from at least three sources: structural uncertainty, performance measure uncertainty, and computational uncertainty.

Structural uncertainty. Structural uncertainty exists because of unexamined models that could also produce relevant predictions. The large set of model candidates I used represents only a subset of the potential model structures that could forecast future fire size. Alternative candidates could be produced in multiple ways: changing the model selection constraints, adding covariates and dimensions, and exploring alternative model structures. I constrained the model selection process so that, at most, one covariate was selected to represent each of seven dimensions of the fire environment. However, this constraint could be loosened to allow more covariates to represent each dimension, which would better represent the environmental processes that modify wildfire size at unique temporal scales, such as drought (Barbero et al. 2015). Additional dimensions could also represent other important factors relevant to fire size, such as suppression levels and fuel availability. Further, different rules that map the covariates

onto each dimension could be adopted, and other covariates could also be included. In our analysis, the fuel moisture content dimension was represented by one of five drought indices, but dozens of other candidates could be included (Zargar et al. 2011). The generalized linear model framework that were used could generate competing model structures within this framework by altering the choice of link functions (Müller and Stadtmüller 2005). Other model structures besides generalized linear models could also be used. For example, the generalized linear model framework (which produces predictions from linear combinations of covariates) could be replaced by a generalized additive model framework (which produces predictions from linear combinations of functions of covariates) (Hastie and Tibshirani 1987). Non-regression methods, such as decision trees, could also produce predictions of final fire size (De'ath and Fabricius 2000, Coffield et al. 2019).

Performance measures. Although the three performance measures examined in this study approximate the preferences of some wildfire managers, other performance measures could be examined instead. For example, the quality of wildfire spread models are often quantified with modeling efficiency, mean absolute error, or mean absolute percent error (Alexander and Cruz 2006) none of which were used in this analysis. User preferences are complex, varying across individual users and contexts, and they need not be represented with a single performance measure. For example one model user might want a model that performs well on average across multiple objectives (Hummel et al. 2013). The preferences of this model user could be represented by a combination of multiple performance measures (e.g., a weighted sum of model rankings calculated with multiple performance measures, with the weights representing user preferences) . Uncertainty in the choice of performance measure is highly relevant because it influences model choice and predictions. In our study, the model that, on average, most closely

predicted fire size was rarely the same as the model that best predicted very large fires when they did occur, or the model that best approximated the distribution of fire size. Predictions of fire size could then differ, sometimes strongly, based on the uncertain choice of performance measure. Given the high uncertainty surrounding the choice of performance measure, multiple model approaches are well-suited for these problems, so I emphasize our recommendation that they be adopted.

Computational uncertainty. Uncertainty also arises when a model is incorrectly identified as "best." In our study, the probability of this error was minimized through the simulation procedures, but other models may also be optimal. Indeed, for a minority of cases (Table 3.5), despite running thousands of simulations, I was unable to identify a statistically best model. Moreover, I observed that models identified as "optimal" by MCCV could perform worse in the holdout analysis than other models that MCCV identified as "suboptimal." Furthermore, suboptimal models may actually be optimal for certain subsets of wildfire events, such as those that occur in particular locations, times of year, or atmospheric conditions. Other resampling techniques could estimate performance measures, which could alter results in ways that may be difficult to anticipate, representing a lingering source of uncertainty.

For each of these three sources of uncertainty, for practical reasons, I had to constrain the possible choices available. I chose fewer model structures and performance measures and evaluated only one potential model candidate. Thus, an apparent conflict exists between the intended purpose of the triple criteria models — to incorporate uncertainty into predictions of future fire size — and its imperfect representation of the full range of uncertainty. The conflict is acceptable, however, because TCMS trades off the completeness of the set of model candidates for computational tractability. Even if we were not constrained by computational limitations, few

would argue that model complexity that would be required to perfectly predict fire size is justified as the level of data inputs required would render the models impractical for decision making applications (Anderson and Burnham 2004). Although the TCMS methodology does not eliminate the problems associated with single-model approaches, it offers a compromise that mitigates these issues.

3.4.5 *Future work*

Previous researchers have recently developed models to predict the size of a wildfire at ignition (Coffield et al. 2019). Our TCMS approach builds on this work in at least three ways. First, distinct classes of models are used for both the first day of a wildfire and later in the wildfire's lifetime. This flexibility in model structures allows for improved prediction of fire size that cannot be achieved with first-day models alone. Indeed I found that the predictive ability of the accumulated evidence models was often better than the first day models. Second, the model uncertainty is incorporated into the TCMS, which cannot be addressed with single models. Third, rather than attempting to predict model preferences, TCMS allows model users to select the most appropriate performance measure for the intended application or, at the very least, assess the sensitivity of predictions to this uncertainty. The approach I used is flexible and can be revised in other contexts to address these uncertainties and improve the quality of fire size predictions.

Intentional reduction of model uncertainties is essential to use TCMS as firefighting decision aides, but more data are needed to accomplish this. In our study, observational holdout data were limited to one year, and some divisions had no fire information at all. Data collection is ongoing to validate estimates of predictive performance, identify models for informing real-time fire decisions, and refine model predictions with bias-correcting techniques and model

averaging (Hoeting et al. 1999). More data are also needed to identify the contexts in which the application of TCMS models is competitive to complex simulators in terms of predictive performance.

The use of statistical modeling to generate predictions of final fire size has the potential to complement information obtained by more complex and computationally expensive wildfire simulators. For instance, the TCMS predictions could work in parallel with output from more complex models to provide a real-time "gut check" of which wildfire characteristics to statistically expect. Hence, the models could be used like wildfire simulators to inform decisions like firefighting strategies, and resource allocation. This is particularly true if covariates like suppression are included in the models to inform of what the likely consequences of a management action would be. These models may also provide a computational shortcut in analyses where output from complex wildfire simulators would be ideal but infeasible. One example of this would be in climate change projections, where generating the complex daily gridded data required to run simulators like FSPro is highly uncertain, whereas the data needed to run the first-day models at least can be collected from climate models. The fact that near-similar performance, at least in terms of fire size, can be generated using drastically simpler models than FSPro, suggests that at least some of the parameters in these and analogous models could be removed with little consequence.

The use of statistical forecasts of fire size represents a potentially powerful tool for decision makers. However, its utility can only be realized if it is understood and applied by relevant decision makers. To enhance the value of these forecasts, further work is needed to evaluate how wildfire managers interpret model uncertainty, identify methods for discouraging

misinterpretation of TCMS output, and provide consultation on contextual information (e.g., fuel, population, nearby fires, and suppression demand).

Although broad themes of covariates were commonly used across the TCMS (e.g. dryness indicators), that different covariates are used when different performance measures are substituted also suggests that types of wildfire events are influenced by different atmospheric processes. For instance, although there is certainly regional variability, it is plausible that very large wildfire behavior may be largely mediated by longer term, severe, drought conditions, whereas the simple rise and fall of average fire size over the year may be governed by short-term atmospheric processes. The consequence of this relationship would be that models that optimize recall, which try to predict extreme events, may prefer the former as a covariate, whereas the short-term atmospheric variables may be preferred if the overall central tendency of the wildfire size distribution is what is prioritized. Future work should examine the effects of performance measure on the observed relationships with atmospherics more closely, as not only do they suggest that the choice of performance measure should be well-defined, but they also might be useful for isolating specific fire-atmosphere processes that are of broader scientific interest.

3.5 CONCLUSIONS

Our analysis provides three major insights relevant to building statistical models of fire size. (1) The predictions of fire size and the structure of models used to generate these predictions can be highly sensitive to the particular needs of model users; therefore these should be identified in advance and inform the model selection procedure. (2) Prediction of final fire size can, in some contexts, be improved by using wildfire growth information that becomes available after the ignition day. The accumulated evidence models that include this information often show better

performance (lower errors, higher recall, and higher likelihood) using fewer parameters than models that predict fire size on ignition day. On-the-ground use of wildfire models could consider an adaptive approach in which one type of model is used early in the fire's life history and a more complex model is built once additional data are available. (3) The simple statistical models presented here produced estimates of fire size that were often similar to FSPro, suggesting that adequate predictions of fire size can often be obtained with relatively few data inputs. Despite the simplicity of the models used here, they show promising predictive ability and potential for use as a decision aid for fire managers.

3.6 REFERENCES

Alexander, M.E., & Cruz, M.G. (2006). Evaluating a model for predicting active crown fire rate of spread using wildfire observations. *Canadian Journal of Forest Research*. *36*(11), 3015-3028

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, *63*.

Barbero, R., Abatzoglou, J.T., Kolden, C.A., Hegewisch, K.C., Larkin, N.K., & Podschwit, H. (2015). Multi□scalar influence of weather and climate on very large□fires in the Eastern United States. *International Journal of Climatology*. 35(8), 2180-2186.

Barbero, R., Abatzoglou, J.T., Steel, E. A., & Larkin, N.K. (2014). Modeling very large-fire occurrences over the continental United States from weather and climate forcing. *Environmental Research Letters*. 9(12), 124009.

Burton, T.A. (2005). Fish and stream habitat risks from uncharacteristic wildfire: observations from 17 years of fire-related disturbances on the Boise National Forest, Idaho. *Forest Ecology and Management*, *211*(1-2), 140-149.

Coffield, S.R., Graff, C.A., Chen, Y., Smyth, P., Foufoula-Georgiou, E., & Randerson, J.T. (2019). Machine learning to predict final fire size at the time of ignition. *International Journal of Wildland Fire.* 28(11), 861–873.

Gebert, Krista M., David E. Calkin, and Yoder, J. (2007) Estimating suppression expenditures for individual large wildland fires. *Western Journal of Applied Forestry.* 22(3), 188-196.

González-Cabán, A. Economic Cost of Initial Attack and Large-Fire Suppression. (1983). USDA Forest Service General Technical Report PSW-068; U.S. Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station: Berkeley, CA, USA, 7p.

De'ath, G., & Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), 3178-3192.

Eidenshink, J., Schwind, B., Brewer, K., Zhu, Z. L., Quayle, B., & Howard, S. (2007). A project for monitoring trends in burn severity. *Fire Ecology*, *3*(1), 3-21.

Filippi, J. B., Mallet, V., & Nader, B. (2014). Evaluation of forest fire models on a large observation database. *Natural Hazards and Earth System Sciences.* European Geosciences Union, 2014, 14, pp.3077 - 3091.

Finney, M.A., Grenfell, I.C., McHugh, C.W., Seli, R. C., Trethewey, D., Stratton, R.D., & Brittain, S. (2011). A method for ensemble wildland fire simulation. *Environmental Modeling & Assessment.* 16(2), 153-167.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371-386.

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science.* 382-401.

Hummel, S., Kennedy, M., & Steel, E.A. (2013). Assessing forest vegetation and fire simulation model performance after the Cold Springs wildfire, Washington USA. *Forest Ecology and Management.* 287, 40-52.

Krueger, E.S., Ochsner, T.E., Engle, D.M., Carlson, J.D., Twidwell, D., & Fuhlendorf, S.D. (2015). Soil moisture affects growing-season wildfire size in the Southern Great Plains. *Soil Science Society of America Journal.* 79(6), 1567-1576.

Littell, J.S., McKenzie, D., Kerns, B.K., Cushman, S., & Shaw, C.G. (2011). Managing uncertainty in climate□driven ecological models to inform adaptation to climate change. *Ecosphere*, *2*(9), 1-19.

Moeltner, K., Kim, M.K., Zhu, E., & Yang, W. (2013). Wildfire smoke and health impacts: A closer look at fire attributes and their marginal effects. *Journal of Environmental Economics and Management.* 66(3), 476-496.

Morgan, M.G., Henrion, M., & Small, M. (1992). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press.

Müller, H.G., & Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33(2), 774-805.

Murnane, R.J. (2006). Catastrophe risk models for wildfires in the wildland–urban interface: What insurers need. *Natural Hazards Review.* 7(4), 150-156.

Papadopoulos, G.D., & Pavlidou, F.N. (2011). A comparative review on wildfire simulators. *IEEE Systems Journal*. 5(2), 233-243.

Picard, R.R., & Cook, R.D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*. 79(387), 575-583.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Riley, S.J., DeGloria, S.D., & Elliot, R. (1999). Index that quantifies topographic heterogeneity. intermountain Journal of sciences, 5(1-4), 23-27.

Rupp, T.S., Olson, M., Adams, L.G., Dale, B.W., Joly, K., Henkelman, J., Collins, W.B. & Starfield, A.M. (2006). Simulating the influences of various fire regimes on caribou winter habitat. *Ecological Applications*. 16(5), 1730-1743.

Sanderson, E.W., Jaiteh, M., Levy, M.A., Redford, K.H., Wannebo, A.V., & Woolmer, G. (2002). The human footprint and the last of the wild: the human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience*. 52(10), 891-904.

Sessions, J., Bettinger, P., Buckman, R., Newton, M., & Hamann, J. (2004). Hastening the return of complex forests following fire: the consequences of delay. *Journal of Forestry*. 102(3), 38-45.

Slocum, M.G., Beckage, B., Platt, W.J., Orzell, S.L., & Taylor, W. (2010). Effect of climate on wildfire size: a cross-scale analysis. *Ecosystems*. *13*(6), 828-840.

Stavros, E.N., Abatzoglou, J., Larkin, N.K., McKenzie, D., & Steel, E.A. (2014). Climate and very large wildland fires in the contiguous western USA. *International Journal of Wildland Fire*. *23*(7), 899-914.

Stephens, S.L., Burrows, N., Buyantuyev, A., Gray, R.W., Keane, R.E., Kubian, R., Liu, S., Seijo, F., Shu, L., Tolhurst, K.G. & Van Wagtendonk, J.W. (2014) Temperate and boreal forest mega-fires: Characteristics and challenges. *Frontiers in Ecology and the Environment.* 12(2), 115–122.

Taylor, S.W., Woolford, D.G., Dean, C.B., & Martell, D.L. (2013). Wildfire prediction to inform management: statistical science challenges. *Statistical Science.* 586-615.

Thies, W.G., Westlind, D.J., Loewen, M., & Brenner, G. (2006). Prediction of delayed mortality of fire-damaged ponderosa pine following prescribed fires in eastern Oregon, USA. *International Journal of Wildland Fire.* 15(1), 19-29.

Weiss, A., & Hays, C.J. (2005). Calculating daily mean air temperatures by different methods: implications from a non-linear algorithm. *Agricultural and Forest Meteorology*, *128*(1-2), 57-65.

Willmott, C.J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society.* 63(11), 1309-1313.

Zargar, A., Sadiq, R., Naser, B., & Khan, F.I. (2011). A review of drought indices. *Environmental Reviews*, *19*(NA), 333-349.

3.7 APPENDIX

Table 3.5. Monte Carlo cross validation (MCCV) convergence statistics for first-day and accumulated evidence model classes. Results are reported for each of the three elements of the triple criteria model set: errors (E), recall (R), likelihood (L).

Biogeographical division	First-day				Accumulated evidence			
	n	E	R	L	n	E	R	L
Hot Continental	5950	***	***	***	7891	***	ns	***
(Sub)Tropical Steppe	5950	***	***	***	7891	***	***	***
Temperate Desert	5950	***	***	***	7891	***	ns	***
Temperate Steppe Mtns	5950	***	***	***	7832	***	***	***
Temperate Desert Mtns.	5950	***	*	***	7832	***	***	***
(Sub)Tropical Desert	5950	***	***	ns	7832	ns	***	***
Mediterranean Mtns.	5950	***	***	***	7729	***	.	***
Mediterranean	5950	***	***	***	7729	.	ns	***
Temperate Steppe	5950	***	***	***	7729	***	***	ns
Marine Forest Mtns	5950	***	***	.	7729	**	***	***
Subtropical	5950	***	***	***	7628	***	***	***
Prairie	5950	***	***	ns	7628	***	***	***
(Sub)Tropical Mtns.	5950	***	***	ns	7628	***	***	***
Warm Continental	2050	***	.	ns	7628	***	NA	ns
Savanna	5950	***	***	ns	7531	ns	***	***

Table 3.6. Monte Carlo cross validation results of first-day models (FDM) and accumulated evidence models (AEM).

Biogeographical division	E		R		L	
	FDM	AEM	FDM	AEM	FDM	AEM
Hot Continental	2.44	1.47	18	99	0.92	0.90
(Sub)Tropical Steppe	2.93	2.25	31	84	1.14	0.82
Temperate Desert	3.37	2.08	38	87	0.83	1.36
Temperate Steppe Mtns	3.65	2.09	39	83	1.42	1.18
Temperate Desert Mtns	2.86	1.84	32	86	1.21	1.08
(Sub)Tropical Desert	2.75	2.34	34	92	1.17	1.08
Mediterranean Mtns.	3.31	2.35	37	86	1.29	1.25
Mediterranean	2.95	2.17	37	81	1.23	0.92
Temperate Steppe	3.01	2.01	34	89	1.25	0.78
Marine Forest Mtns	3.41	2.18	40	75	1.43	1.31
Subtropical	2.70	2.17	21	83	0.91	0.98
Prairie	2.35	1.32	25	89	1.09	0.37
(Sub)Tropical Mtns.	3.14	2.08	38	84	1.29	1.25
Warm Continental	2.72	1.55	19	67	0.95	2.58
Savanna	3.35	1.63	35	99	1.32	1.05

Table 3.7. Holdout validation results of first-day models (FDM) and accumulated evidence models (AEM).

Biogeographical division	E		R		L	
	FDM	AEM	FDM	AEM	FDM	AEM
(Sub)Tropical Steppe	1.91	2.23	23	65	1.03	1.13
Temperate Desert	11.14	6.01	31	91	2.44	3.58
Temperate Steppe Mtns	3.77	3.65	37	69	1.56	1.18
Temperate Desert Mtns	2.67	1.97	31	99	1.55	1.05
(Sub)Tropical Desert	2.13	1.84	NA	NA	1.01	0.96
Mediterranean Mtns	13.72	10.91	31	82	2.74	1.93
Mediterranean	2.13	2.01	NA	NA	0.43	0.60
Marine Forest Mtns	5.97	6.49	43	78	2.17	1.41
Prairie	1.8	1.58	NA	NA	1.36	0.11
(Sub)Tropical Mtns	11.3	8.31	31	83	1.82	2.05
Savanna	2.28	3.2	34	99	2.08	0.97

Table 3.8. Elements of triple criteria model set in first-day model class.

Biogeographical division	Performance measure	Formula (Inverse link in italics)
Hot Continental	Error	*0.948112-0.004078×bi+0.018616×hii*
	Recall	*1.27934-0.02555×fm100-0.03978×vs+0.02111×hii*
	Likelihood	*0.77860+ 0.02052×hii*
(Sub)Tropical Steppe	Error	0.002330 -23.158800×sph -0.029612×fm100+ 0.090204×vs+0.001399×tri
	Recall	-3.4591- 42.5347×sph+0.002638×erc+0.102465×vs+0.011116 ×temp+0.001375×tri-0.020160×hii
	Likelihood	-5.003003- 56.871438×sph+0.096497×vs+0.016743×temp+0.001518 ×tri
Temperate Desert	Error	*7.07518+19.46829×sph-0.01777×pet-0.02208×temp+0.01730×hii*
	Recall	-7.06532 -0.03377 ×fm100+0.02966×vs+0.02628×temp-0.02148×hii
	Likelihood	-7.06532 -0.03377 ×fm100+0.02966×vs+0.02628×temp-0.02148×hii
Temperate Steppe Mtns	Error	3.7026262+0.1528351×pet-0.0451129×vs-0.0139327×temp+0.0007522×tri-0.0158561×hii
	Recall	-0.08153-22.3452×sph+0.1193×pet-0.0303×vs+0.0007×tri-0.0166×hii
	Likelihood	-0.2593726 +0.1089096 ×pet+0.0007774 ×tri-0.0171863×hii
Temperate Desert Mtns	Error	0.59618+71.76390×sph-0.09327×fm1000
	Recall	0.59618+71.76390×sph-0.09327×fm1000
	Likelihood	0.59618+71.76390×sph-0.09327×fm1000
(Sub)Tropical Desert	Error	-0.02887-62.31×sph+0.09339×pet+0.0009×tri-0.03048×hii
	Recall	-6.3344- 70.1236×sph+0.0042×erc+0.0508×vs+0.0220×temp+0.0009×tri-0.0314×hii
	Likelihood	*3.3543642 +66.0822775×sph+0.0325675×fm1000-0.0109373 ×temp-0.0004865×tri+0.0266402×hii*
Mediterranean Mtns	Error	-26.7-108.7×sph-0.1727×pet+0.1685×vs+0.09562×temp+0.0002669×elev-0.02163×hii
	Recall	-21.8599-107.8704×sph-0.1476×pet+0.1363×vs+0.0794×temp+0.0009×tri-0.0247×hii
	Likelihood	-26.7-108.7×sph-0.1727×pet+0.1685×vs+0.09562×temp+0.0002669×elev-0.02163×hii
Mediterranean	Error	-0.630783+0.008583×erc+0.002778×tri
	Recall	0.7613-104.4512×sph+0.0025×tri-0.0078×hii
	Likelihood	0.434711-89.183582×sph+0.001126×elev
Temperate Steppe	Error	7.530634 +0.307820×vpd-0.036627×fm1000+0.033162×vs-0.025814×temp+0.001595×tri-0.016762×hii

	Recall	-7.4169-74.094022×sph-0.026851×fm1000+0.027899×vs+0.028573×temp+0.001503×tri-0.016758×hii
	Likelihood	-6.77883 -74.72404 ×sph-0.02538 ×fm1000+0.02680×temp+0.00137 ×tri-0.01673×hii
Marine Forest Mtns	Error	-8.5134492 -0.0596503×fm100+0.0306299×temp+0.0004133×elev
	Recall	-9.435-0.0804586×fm1000+0.0351861×temp +0.0004416×elev-0.0088554×hii
	Likelihood	0.2464093 +0.2242725×vpd-0.0701486×fm1000+0.0004781×elev
Subtropical	Error	*0.4736-31.084207×sph+0.051868×fm1000+0.008344×tri*
	Recall	*5.843723+0.035661×fm1000-0.019801×temp+0.008294×tri+0.026121×hii*
	Likelihood	*1.5833 -0.103449×pet+0.007781×tri*
Prairie	Error	-0.498810+0.005281×bi+0.001247×elev
	Recall	*8.8414070+62.7625505×sph-0.0176694×fm100-0.0275291×temp-0.0008669×elev+0.0084906×hii*
	Likelihood	*1.4571-0.005037 ×bi-0.001080×elev*
(Sub)Tropical Mtns	Error	*0.9840999 +35.7303×sph+0.0323553×fm100-0.0003364×elev*
	Recall	*0.8908604 +38.5346484×sph+0.0318895×fm1000-0.0003219×elev+0.0085985×hii*
	Likelihood	*0.9840999 +35.7303×sph+0.0323553×fm100-0.0003364×elev*
Warm Continental	Error	-7.57938+0.08049×fm100+0.02291×temp
	Recall	-17.1531-0.6163×vpd+0.08298×fm1000+0.05795×temp-0.01188×tri
	Likelihood	*1.103*
Savanna	Error	2.76371 -0.09686×fm1000-0.12257×vs-0.02663×hii
	Recall	3.2427-0.1065×fm100-0.1418×vs-0.1469*elev-0.0298×hii
	Likelihood	*-0.29888 +32.36797 ×sph+0.11159*×vs+*0.03169×hii*

Table 3.9. Elements of triple criteria model set in accumulated evidence model class.

Biogeographical division	Measure	Method	Formula (Italicized: Inverse link)
Hot Continental	Error	A	1.9848+0.2867×ge.25c
	Recall	M	-9.11404-0.26292×vs+0.03264×temp+0.06175×hii+0.43280×ge.25c
	Likelihood	A	*0.13741-0.03425×ge.25c*
(Sub)Tropical Steppe	Error	A	2.0550-3.2596×sph+0.2048×ge.25c
	Recall	A	2.0349 + 0.2088×ge.25c
	Likelihood	M	-1.399075 +0.155866×pet+0.003305×tri-0.300065×ge.10c
Temperate Desert	Error	A	1.9888496+0.0172920×pet-0.0001001×tri+0.2659163×ge.100c
	Recall	A	2.792+0.01225×pet-0.0109×vs-0.00268×temp-0.000068×tri-0.00083×hii+0.21364×ge.25c
	Likelihood	M	-0.2957731+0.0003473×elev-0.4473349 ×ge.10c
Temperate Steppe Mtns	Error	A	*0.1255205 -0.0238465×ge.25c+0.0001886×hii*
	Recall	A	1.44857664-7.06417287×sph+0.00561907×fm100+0.00200392×temp+0.00001221×elev+0.21066088×ge.25c
	Likelihood	A	*0.1255205 -0.0238465×ge.25c+0.0001886×hii*
Temperate Desert Mtns	Error	A	*0.110307 +0.002465×pet-0.017847×ge.25c*
	Recall	M	0.8502-1.5067×ge.25c
	Likelihood	A	*0.110307 +0.002465×pet-0.017847×ge.25c*
(Sub)Tropical Desert	Error	A	2.17477-0.01035×fm1000+0.26077×ge.100c
	Recall	M	0.06469-0.05161×ge.25c
	Likelihood	M	*0.8014+0.2233×ge.10c*
Mediterranean Mtns	Error	A	*0.122878+0.001128×fm100-0.026515×ge.25c*
	Recall	A	2.105535971-0.008467787×fm100-0.000009669×elev+0.226469345×ge.25c
	Likelihood	A	*0.122878+0.001128×fm100-0.026515×ge.25c*
Mediterranean	Error	M	*0.8211+1.0652×ge.25c*
	Recall	M	-10.85211-0.20657×pet +0.37026×vs+0.03906×temp - 1.13391×ge.25c-0.02012×hii
	Likelihood	M	*0.8211+1.0652×ge.25c*
Temperate Steppe	Error	A	*0.1155865+0.0015903×fm100-0.0227096×ge.25c+0.0002991×hii*
	Recall	M	-0.211181+0.002935×bi+0.005691×tri-0.354341×ge.25c-0.043752×hii
	Likelihood	M	-0.483999+0.004684×tri-0.198361×ge.25c
Marine Forest Mtns	Error	A	*0.12560-0.02131×ge.25c*
	Recall	M	1.162-0.978×ge.10c
	Likelihood	A	2.0746+0.1859×ge.25c
Subtropical	Error	A	2.055442 +0.000506×elev+0.286748×ge.100c
	Recall	A	1.1127521-6.8306476×sph-0.0007494×bi+0.0036409×temp+0.1886660×ge.25c-0.0031733×hii
	Likelihood	M	*0.5908 +0.7658×ge.10c*
Prairie	Error	M	-0.677022 +0.009634×bi-0.666137×ge.10c
	Recall	M	-0.72846 +0.14383×vpd+0.00634×bi-0.76784×ge.10c
	Likelihood	M	-0.2081-0.7200×ge.10c
(Sub)Tropical Mtns	Error	A	*0.1479608-0.0003056×erc-0.0203260×ge.25c*
	Recall	M	*0.31388-0.02171×vs+0.52779×ge.10c*
	Likelihood	A	*0.1479608-0.0003056×erc-0.0203260×ge.25c*
Warm Continental	Error	A	1.9947 +0.3315×ge.25c
	Recall	M	1.05550 -0.49605 ×ge.10c-0.03085×hii

	Likelihood	M	0.6880 -0.4393×ge.10c
Savanna	Error	A	2.0389+ 0.2358×ge.25c
	Recall	A	1.984475+0.003192×erc+0.208951×ge.25c
	Likelihood	A	*0.13017-0.02734×ge.25c*

Table 3.10. Comparison of mean multiplicative error estimates of simple triple criteria model set elements (first-day models, accumulated evidence models), and complex wildfire simulators (FSPro).

Biogeographical division	Triple Criteria Model Set		Wildfire Simulators
	First-day model	Accumulated evidence model	FSPro [11]
(Sub)Tropical Steppe	1.91	2.23	1.79
Temperate Desert	11.14	7.25	6.11
Temperate Steppe Mtns	3.77	3.65	5.08
Mediterranean Mtns	13.72	10.91	4.18
Marine Forest Mtns	5.97	6.49	1.78
(Sub)Tropical Mtns	11.30	8.31	2.73

Chapter 4. A METAMODEL FOR INTERGRATING MODEL CHOICE UNCERTAINTIES AND UTILITY FUNCTION UNCERTAINTIES INTO FIRE SIZE FORECASTS

4.1 INTRODUCTION

There are a variety of situations where there is high demand for predictions of fire size. Many decisionmakers and lay audiences monitor the progress of nearby wildfires for reasons ranging from curiosity to concerns about material values threatened by wildfire growth. Decision makers that could utilize predictions of final fire size include public health professionals (Moeltner et al. 2013), public safety officials (Intini et al. 2019), and the general public (Taylor et al. 2005). Where fire size correlates with high severity fire (Miller et al. 2008), land managers could use such predictions to develop post-fire restoration strategies and priorities. Once wildfires are discovered, estimates of final fire size are often used by (Thompson et al. 2017) and requested from fire managers (National Interagency Fire Center 2014). Useful predictions of final wildfire size are, however, difficult because of the numerous uncertainties associated with quantitative forecasts and because of uncertainties in how end-users apply the quantitative forecasts.

Uncertainties in quantitative forecasts are near-ubiquitous and can arise in the context of wildfire prediction from multiple sources. For instance, random variation is inherent to most natural phenomenon. Hence, even if the probability distribution describing the population of interest were known with certainty, predicting any one realization of the probability distribution would always contain error. The challenges associated with predicting any one realization of a probability distribution are further compounded by uncertainties arising from inevitable measurement errors, both random and systematic (Morgan and Henrion 1990, Uusitalo et al.

2015). This tendency is also seen within the context of wildfire prediction, as systematic errors are common in satellite derived estimates of burned area, particularly in topographically complex areas (Kolden and Weisberg 2007), as well as disagreements across sources in the estimates of daily fire size (Podschwit et al. 2020b). The sources of uncertainty mentioned so far have assumed that the model describing the population of interest is known with certainty, which is seldom the case. Much more typically, there are multiple competing models to describe the natural phenomena that can produce predictions that disagree (Morgan and Henrion 1990). For instance, wildfire size is controlled by multiple factors that work at varying time scales. These factors include weather and climate, topography, fire suppression, wildfire growth, vegetation effects, anthropogenic factors, and past wildfire (Cui and Perera 2008). Building a model of wildfire size not only requires selecting which of these factors are relevant and estimating the effect of that factor on wildfire size, but also selecting which variable best represents them. For example even though it is well known that the dryness of environment is an important mediator of wildfire activity (Meyn et al. 2007), there are several drought indexes that could be used as a proxy for dryness (Heim 2002, Cullen et al. 2020). Moreover, the importance of certain factors may vary with time. Although certain environmental factors may be highly relevant to predicting wildfire size in the first-days of a wildfire, near the end of a wildfire's lifetime other factors may be more useful. For instance, the current size of the fire could eventually provide near-perfect predictions of the final fire size when little additional growth is expected. As seen in the previous section, more useful, albeit slightly less accurate predictions can be produced using daily fire size increments as covariates. However, knowing when useful predictions can be derived from current size and fire growth information is not always obvious.

A further reason useful predictions of wildfire size are difficult to develop is not associated with the numerous forms of model uncertainty, but rather with unknowns regarding how the information will be used by end-users. For instance, there is often uncertainty regarding how model quality should be measured, and multiple utility functions could be selected to identify which models provide "good" predictions (Morgan and Henrion 1990). This uncertainty in the choice of utility function was explored in the context of wildfire size forecasts by (Podschwit et al. 2020a) who identified multiple objectives model users may have. In particular, one consumer of wildfire size forecasts may desire that predictions are close to the truth in the long run; another may desire forecasts that well approximate the entire distribution of fire size; yet another may want a forecast that correctly predicts very-large wildfires when they occur. Even when the utility function is clearly defined, probabilistic forecasts can be used by end-users in multiple ways. A single best, point, estimate may be sufficient for low-stakes users that are not severely impacted if forecasts strongly deviate from the truth. In other cases, there are specific event-spaces or thresholds which if exceeded, are significant to the end-user (Raftery 2016). In this case, the probabilities of those exceedance events may be what is most useful to the end-users. Yet other users may desire an expected point estimate like the low-stakes user but will be severely impacted if the forecast strongly deviates from the truth. Such users may instead benefit from interval estimates that contain the truth with a certain probability – with 80% coverage probabilities being identified as being narrow enough to provide information that is actionable to many decision makers (Raftery 2016). It is also uncertain as to what quantity is relevant to model users. For instance, although fire size may be commonly modeled, the number of simultaneous wildfires within an administrative region may be more relevant to wildland fire fighter operations (Podschwit and Cullen 2020).

While the uncertainty associated with the randomness of the probability distributions that represent a natural phenomenon can be easily simulated, accounting for model uncertainties requires more effort on part of the analyst. One method of doing so is through the use of Bayesian model averaging to create a model of models, hereafter metamodel. The metamodel represents uncertainty about the model structure with weights, which can be used to combine predictions from multiple models into a single probability distribution, hereafter the posterior. Multiple kinds of test statistics relevant to model users – expected values, exceedance probabilities, intervals – can be directly calculated from the posterior and compared to real data to validate the predictive ability of the metamodel (Fragosa et al. 2017). Hence, Bayesian model averaging can at least partially, address the uncertainties present in statistical forecasts of fire size, and may also gauge the usefulness of the metamodel to produce certain test statistics.

Although Bayesian model averaging has been used to explore the effects of model uncertainties in models of very-large fire occurrence (Podschwit et al. 2018b), and the time-varying uncertainties in forecast have been explored without Bayesian Model Averaging by (Podschwit et al. 2020a), there has been no research to date that have combined the time-varying model uncertainties associated with predicting final fire size into a single metamodel. To that end, in this chapter I will use Bayesian Model Averaging to combine existing statistical models of fire size using a metamodel that accounts for both statistical randomness, model uncertainties, and uncertainty in the utility function. The model will be described generally and then applied to a specific example data for the Temperate Steppe Regime Mountains ecoregion. Multiple test statistics are used to determine which contexts the time-varying model averaging (TVMA) metamodel produces useful predictions. In section 2, I will describe the metamodel generally, the methods used to produce and combine the individual models to predict fire size. I will also

describe the methods used to fit the TVMA metamodel to specific case data in the Rocky Mountains, as well as the methods used to validate it. In section 3, I report details regarding the predictive performance of the TVMA metamodel in the Rocky Mountains as compared to the individual constituent models. In section 4, I will discuss the benefits of the TVMA metamodel described in section 2, some of their limitations, and implications to end-users. In section 5, I will summarize the overall conclusions that are warranted from this analysis.

4.2 METHODS

4.2.1 *Overview*

The metamodel consists of a time-varying weighted average of three model classes, each intended for use in specific stage of the wildfire's lifetime. The first-day models (FDMs) use information on the first-day of a wildfire to generate predictions, the accumulated evidence models (AEMs) use growth information later in the fire's lifetime to generate more refined predictions, and the end-of-fire models (EOFMs) assume simple temporal autocorrelations in daily size to predict fire size. The metamodel applies time-constant weights to individual models within each class, and then weighs the average predictions from each class using a discrete-time Markov chain.

4.2.2 *Individual models*

The FDMs and AEMs predict the final fire size, Y, using generalized linear models that use meteorological and topographic covariates from the gridMET data project[1]2, and

2 http://www.climatologylab.org/gridmet.html

anthropogenic covariates in the form of the human influence index from the NASA Socioeconomic Data and Applications Center (Sanderson et al. 2002). A total of nine meteorological covariates were considered: 100-hour fuel moisture, 1000-hour fuel moisture, the model-G energy release component, burning index, daily average temperature, specific humidity, potential evapotranspiration, vapor pressure deficit, and wind speed. Daily average temperature was calculated as one-half the sum of the daily maximum and minimum values (Weiss and Hays 2005). Two topographic variables were considered: elevation and topographic roughness index (Riley et al. 1999). The AEMs included a dummy variable that equals one if the largest growth event exceeded a predefined size threshold and was otherwise zero. The predefined size thresholds were 405 ha, 1012 ha, or 4050 ha of daily burned area. Exploratory analysis has shown that using a continuous version of this dummy variable – the largest daily growth increment – can produce reasonable predictions of subsequent area burned and by extension final fire size. Moreover, this variable has a plausible explanation for its relationship to final fire size, with large daily changes in fire size increasingly the management complexity to the point that it became resistant to suppression and by extension grow larger than slower growing events. However, the disadvantage of that approach is that it is impossible to know if the largest daily growth increment is being realized – what if a future growth increment is larger – whereas the dummy variable does not suffer from this problem. The dummy variable will still report one even if a larger growth event is realized. All the generalized linear models used a Gamma probability distribution with either a log or inverse link function. The AEMs either adopted an additive or geometric method. The additive method predicted the amount of additional area that will burn after the forecast and added this to the current size. The geometric method predicted

the multiplicative factor between the current size and the final size and multiplied this to the current size. A detailed description of these models can be found from Podschwit et al. (2020a).

The set of FDMs and AEMs used by the metamodel are identified using cross-validation results from (Podschwit et al. 2020a). Specifically, the predictive ability - in terms of expected error, likelihood, and recall - of an initial suite models are measured, and the quality each of the models from each class is ranked according to the three performance measures. A final model quality score is produced using a weighted average of the three model ranks. Variability in the weights represent differences in end-user preferences. To simulate these differences and uncertainty in end-user preferences, a sample of 1,000,000 random weights are used to produce the model quality scores and the set of models that are associated with these scores are selected for use in the final ensemble.

In addition to the FDMs and AEMs, the EOFMs are used during the final days of the wildfire when little growth is expected. Two EOFMs are considered, which like the AEMs use information calculated from the burned area time series, X_t, to estimate final wildfire size. The first EOFM predicts that the final wildfire size will equal the current size (Equation 4.1). The second predicts that the final wildfire size will equal the current wildfire size plus yesterday's growth (Equation 4.2).

$$Y \sim \text{LogNormal}(\ln(X_t), \ln(1.0001)); \qquad (4.1)$$

$$Y \sim \text{Normal}(2X_t - X_{t-1}, {}^{1}/_{t+1} + X_t - X_{t-1}); \qquad (4.2)$$

Note that $t \in \{0,1,\ldots,T\}$ and the EOFMs do not produce predictions until $t > 0$.

4.2.3 *Metamodel*

The probability distribution of wildfire size is produced using a time-varying mixture model. Specifically, for each day of a wildfire's lifetime, a probability distribution is produced by randomly sampling the individual models according to time-varying weights. The weights applied *within* each class do not vary with time, and the weights applied *to* each class changes in time according to a discrete-time Markov Chain.

As in any fully Bayesian approach, we must define probability distributions to represent our a priori belief in the model parameters. An uninformative Dirichlet prior is used to produce the within-class weights equations (Equation 4.3). Informative priors are used (Equation 4.4 and 4.5) to represent the elements of the stochastic matrix, A, (Equation 4.6), which is in turn used to generate the time-varying weight applied to model averages within each class is described in (Equation 4.3-4.5).

$$q_{fdm,*}, q_{aem,*}, q_{eofm,*} \sim \text{Dirichlet}(\vec{1}); \quad (4.3)$$

$$p_{1,*} \sim \text{Dirichlet}(\langle 0.9, 0.08, 0.02 \rangle); \quad (4.4)$$

$$p_{2,*} \sim \text{Dirichlet}(\langle 0, 0.9, 0.1 \rangle); \quad (4.5)$$

$$A = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ 0 & p_{22} & p_{23} \\ 0 & 0 & 1 \end{bmatrix}. \quad (4.6)$$

The weights applied to each individual model will vary with time and model class. Specifically, for a wildfire that T days long, the weight, applied to the individual FDMs, AEMs, and EOFMs on day $t \in \{0,1,\dots,T\}$, w_{it}, follows (Equation 4.8).

$$B^t = (1,0,0) \times A^t; \quad (4.7)$$

$$w_{it} = \begin{cases} q_{fdm} \times B_1^t, \text{ If } i \text{ is a FDM} \\ q_{aem} \times B_2^t, \text{ If } i \text{ is an AEM} \\ q_{eofm} \times B_3^t, \text{ If } i \text{ is an EOFM} \end{cases}. \quad (4.8)$$

The within-class weights and transition probabilities are estimated using daily wildfire growth data collected from the ICS209-PLUS dataset (Denis et al. 2020). To avoid duplicative use of the data that were used to train the predictive models, only data from the years 1999-2001 and 2014 are used. The metamodel parameters are fit using JAGS software in the R Programming Language (R Core Team 2020). An initial Markov chain Monte Carlo (MCMC) was run using three parallel chains with a nominal sample size of 5000, thinning interval of 100, burn-in period of 5,000 and adaptive phase of 5,000. Convergence is assessed visually and by using the potential scale reduction factor applied to the central 90th percent of the simulated marginal posterior distributions as a convergence diagnostic (Brooks and Gelman 1998, Podschwit et al. 2018a).

4.2.4 *Validation*

Simulation is used to approximate the daily probability distributions of wildfire size that incorporate three sources of uncertainty: parameter uncertainty, model uncertainty, and statistical uncertainty. Uncertainty in the metamodel parameters – the model weights - is incorporated by randomly selecting weights from the posterior sample. Next, model uncertainty is incorporated by using the weights selected from the previous step to randomly select a model. Finally, statistical uncertainty is incorporated by randomly generating predictions from the selected model. This process is repeated to produce a sample of 10,000 wildfire size predictions for each

day of the wildfire. This posterior sample is then used to assess the performance of predictions produced from the individual models from the three classes compared to those produced from the time-varying model-averaging technique.

Burned area time series data from (Podschwit et al. 2020b) for the years 2018-19 are used to validate the predictive performance. The burned area time series are cleaned by averaging the daily size estimates from all available sources for each day, using linear interpolation to fill-in the missing days, and applying a backwards smoothing pass (Denis et al. 2020). This produces a daily burned area time series with no missing data and no negative growth.

In practice, a lower-bound can be placed on the samples of daily wildfire size predictions since it is certain that a wildfire will be at least as large as the previous day's size. Hence, the samples of daily wildfire size predictions are lower-censored using the burned area time series data so that no predictions are smaller than the previous day's wildfire cumulative area. Values smaller than the previous day's wildfire size are set equal to the daily lower size threshold. The lower-censored sample of daily wildfire size predictions – for both the individual models and the time-varying model averages - are compared to the cleaned burned area series using four methods. First, the typical difference between median daily predictions and the reported final wildfire size are estimated for each day. Second, the probability that the central 80^{th} percentile of predictions contains the reported final wildfire size are estimated for each day. Third, the estimated probabilities of wildfire exceeding 4047 hectares, hereafter a very-large fire (VLF), are estimated for each day for two groups of wildfires: those larger than 4047 hectares and those smaller than 4047 hectares. Lastly, the interquartile range of predictions are estimated for each day.

4.2.5 *Study area and data*

As mentioned in the previous section, burn area data from the ICS209-PLUS dataset (Denis et al. 2019) and WOMBATS dataset (Podschwit et al. 2020b) are used for fitting and validating the metamodel respectively. Both data are available nationally and must be filtered to include only wildfires that burned in the relevant geographic area. In this analysis, I will fit the metamodel for the Temperate Steppe Regime Mountains as defined from the Baileys biogeographical divisions (Bailey 2016). A map of the region, along with the relevant ICS-209 and WOMBATS data are shown in (Figure 4.1).

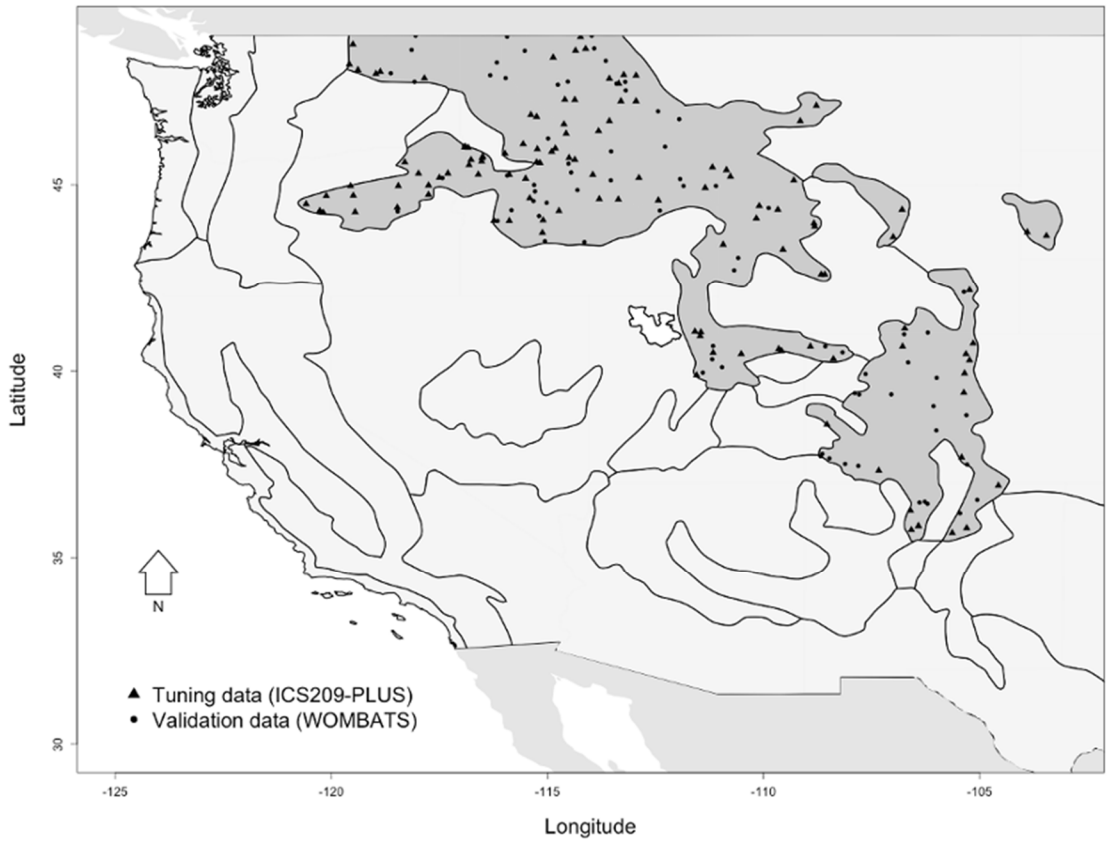


Figure 4.1. A map of the Bailey's divisions, along with the wildfire locations in the Temperate Steppe Regime Mountains that are used in the tuning and validating phases of the analysis. The metamodel parameters are fit using ICS209-PLUS data for the years 1999-2001 and 2014

(n=118). The metamodel predictions are validated using WOMBATS data for the years 2018-2019 (n=75).

4.3 RESULTS

4.3.1 *Model ensemble*

The models used in the FDM and AEM ensembles are shown graphically using a barycentric plot in Figure 4.2. Randomly selected preferences can be produced by randomly selecting a point on the barycentric plot figure (Figure 4.2) and we will report the probabilities that each model in the ensemble would be selected under such a sampling scheme. For the FDMs, a randomly selected preference is likely to yield distinct models. Assuming a uniform distribution of users across the preference space, a plurality (~37.1 percent) of end-users would prefer a model based on potential evapotranspiration, wind speed, temperature, topographic roughness, and the human influence index. However, nearly as many end-users (~36.8 percent) would prefer a simpler version of this model that omits temperature. A slightly smaller faction (~24.4 percent) of end-users would prefer that the model substitutes specific humidity for potential evapotranspiration. End-users who selected models based mostly on likelihood measures would represent the smallest faction (~1.7 percent), and would prefer a model based on potential evapotranspiration, topographic roughness, and the human influence index. However, unlike the FDMs, the choice of model in the AEMs is robust to changes in user preferences. Assuming a uniform distribution of end-users across the preference space, a clear majority of end-users (~84.7 percent) would prefer a model that uses no weather covariates, but uses the human influence index, and a 1012-hectare daily growth increment dummy variable. The remainder of the preference space (~ 15.2 percent) was nearly entirely represented by a model

that included a wind speed variable. The remaining two AEMs would be preferred by less than one percent of end-users assuming they are uniformly distributed about the preference space.

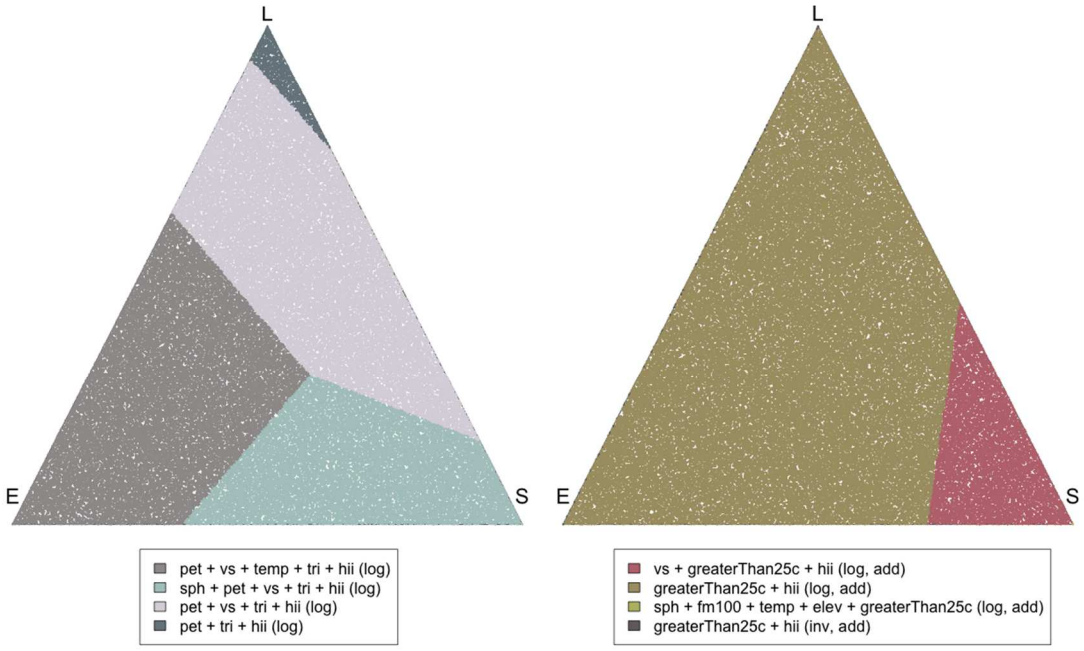


Figure 4.2. Graphical summary of the selection of first-day models and accumulated evidence models. Barycentric plots are used to represent end-user preferences. End-user that gauges model quality using expected errors (E), likelihood (L), or sensitivity (S) entirely, are represented at the vertices of the plot. End-users that that gauge model quality using multiple preferences simultaneously are represented in the interior of the triangle. All models in the preference space are selected for use in the metamodel. For both model classes, the link function is identified in the parenthesis. For AEMs only the method of calculating the final burned area is also identified in the parenthesis. Potential evapotranspiration, pet; specific humidity sph; temperature, temp; wind speed, vs; topographic roughness index, tri; elevation, elev; human influence index, hii

4.3.2 *MCMC diagnostics*

Convergence of the metamodel parameters as estimated from MCMC appeared adequate both from visual inspection of the traceplots, as well as the PSRF diagnostic statistics. Visual

inspection did not reveal any overly-strong autocorrelation and the MCMC chains had a distributions that were in approximate agreement with one another. This was confirmed by the PSRF statistics which was, at worst, only 1.008 when estimating the weights for the AEMs. This worst-case PSRF was below the guidelines identified by (Brooks and Gelman 1998). The PSRF statistics and metamodel parameter estimates are described in Figure 4.3.

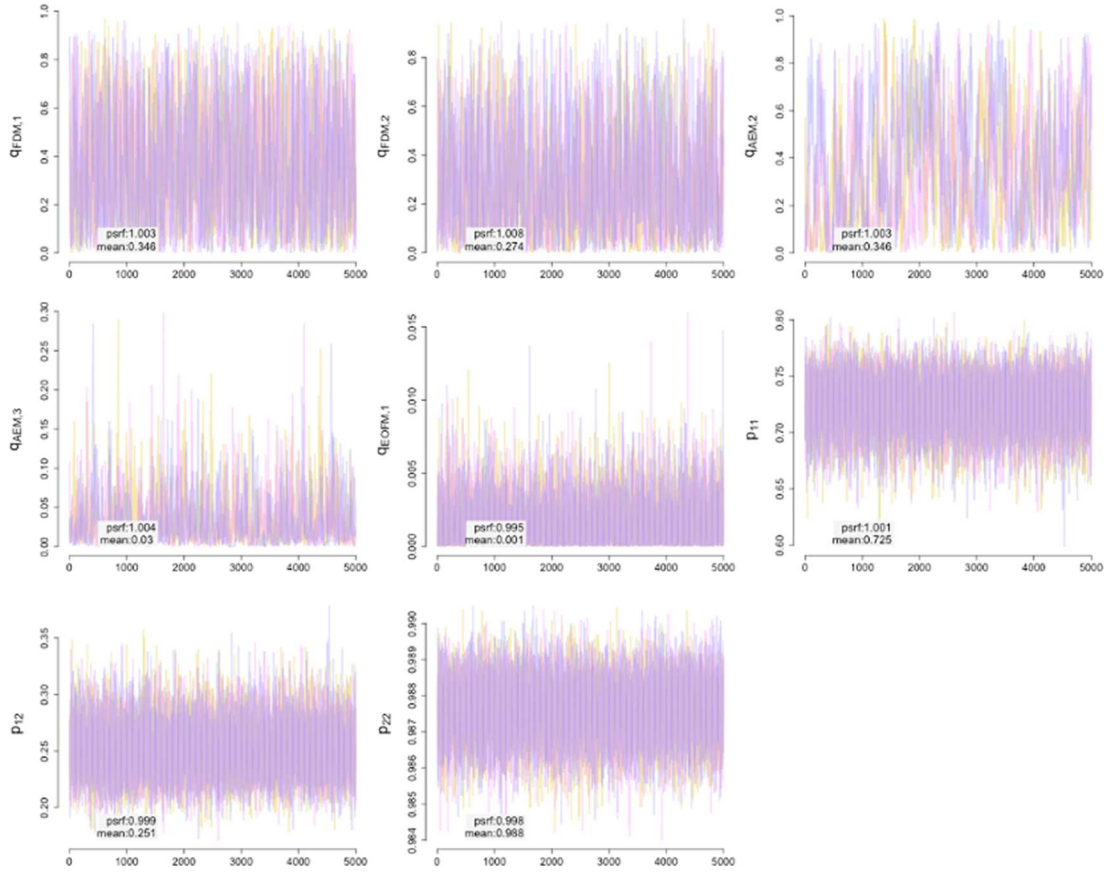


Figure 4.3. Traceplots of posterior from Markov Chain Monte Carlo that graphically summarizes the convergence of the simulations. The results of three simulations are overlaid in each panel and suggest that each realization of the stochastic algorithm is similar. Potential score reduction factors (PSRFs) and parameter estimates are reported in the bottom left of each panel.

4.3.3 *Validation*

An example of the predictions produced from the metamodel is shown in (Figure 4.4). In this example, note that the 80 percent predictive interval is quite wide in the early days of the wildfire, with the lower and upper bound of the predictive interval differing by a factor of more than 16 on the first day. The variability in the 80th percent predictive interval drops quickly during the first week of the wildfire and continues shrinking over the wildfire's lifetime. At the end of the first week, the upper and lower bound differ by a factor of 5, and by the end of the second week a factor of 4, by the end of the third week a factor of 2. All daily predictive intervals contained the final fire size in this example. The median estimate from each day was variable in the first three weeks of the wildfire but changed little afterward once the wildfire had finished growing. I hereafter consider the typical results observed across all 75 wildfires in the validation data.

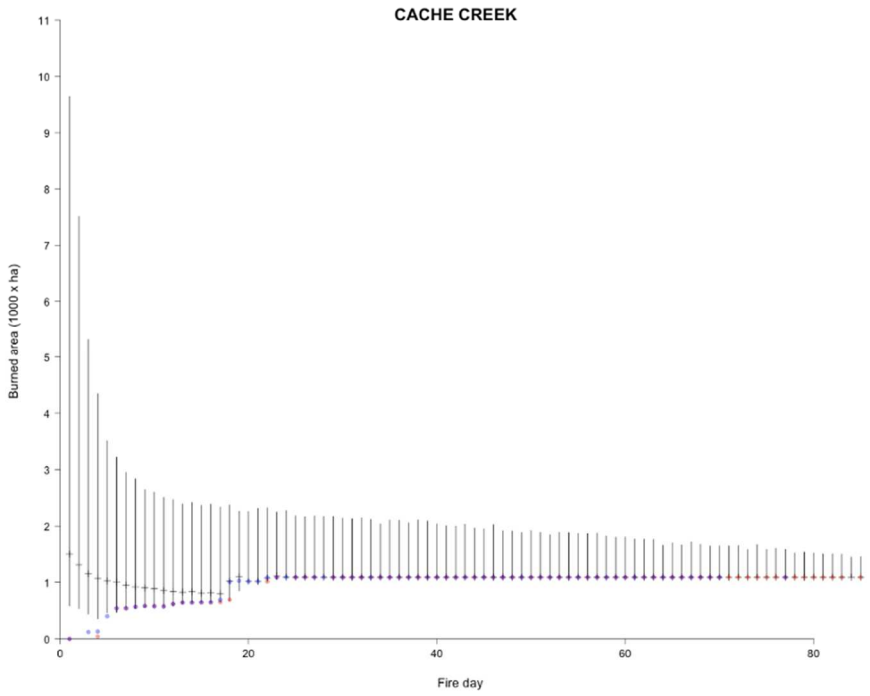


Figure 4.4. The central 80th percentile of predictions of final fire size produced from TVMA forecast for the Cache Creek wildfire (2018, Colorado), which is selected as an example of a typical forecast because the percent errors of the forecast are close to the validation set's mean. The daily median posterior prediction is shown with a horizontal line, and the data from WOMBATS data is shown as colored points: InciWeb (blue) and Incident Management Situation Reports (red).

The average percent difference between the median TVMA fire size prediction and the reported fire size was fairly small and did not change drastically over time. Specifically, the TVMA predictions underestimate the reported final fire size by about 3.5 percent. The largest errors tended to occur about four days after ignition, at which time the TVMA predictions underestimate the reported final fire size by about 10.5 percent. After this peak, the percent error in TVMA predictions typically fell between 1.3 percent (Q_1) and 5.3 percent (Q_3) until day 60. The average percent difference between the FDM and AEM models were initially approximately similar to that of the TVMAs – within $\pm$ 2 of the TVMA forecast - but fell over time. This eventually produced substantial overestimates of final fire size, by as much as 40 percent in the

former and 32 percent in the latter. The EOFM initially severely underestimated the final fire size but became increasingly accurate as time elapsed from the ignition date (Figure 4.5).

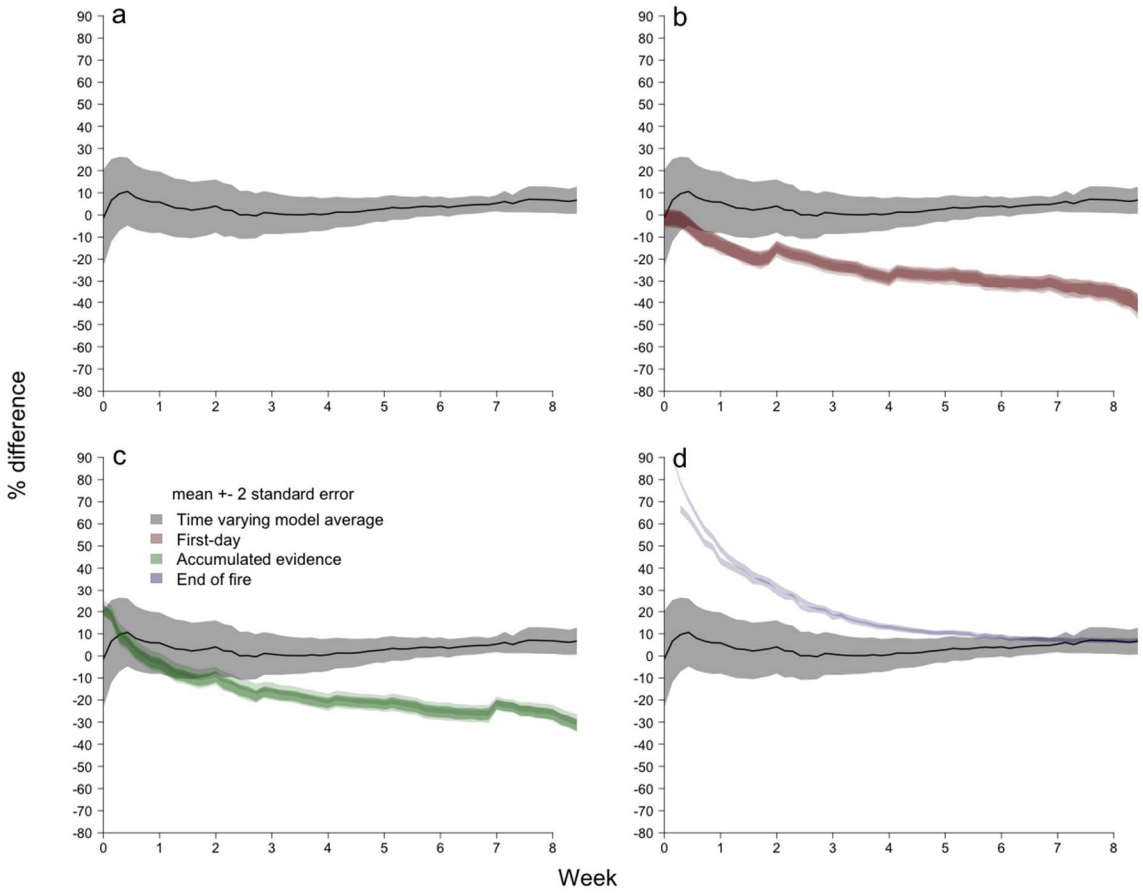


Figure 4.5. The percent difference between the predicted final fire size and the reported final fire size over time for the time varying model average (a), first-day models (b), accumulated evidence models (c), and end-of-fire models (d). Time is shown on the x-axis and is relativized for each for each fire so that values near zero represent how close first day forecasts are. The TVMA forecasts were relatively accurate, typically within 10 percent of the reported size, but showed a slight tendency to underestimate the reported fire size. With few exceptions, the TVMA forecasts tended to show consistently better predictive performance than any of the individual constituent models.

The probability that the central 80 percent of the TVMA posterior contained the final fire size, the coverage probability, was about 76 percent and remained roughly constant over the first

60 days. The coverage probabilities of the FDMs were also roughly constant over time and contained the final fire size between an average of 67 and 71 percent of the time over the first 60 days depending on the individual model. The coverage probabilities of the AEMs were comparable with TVMAs and showed a slight increasing temporal trends. Specifically, the AEMs coverage probabilities of the individual AEMs were slightly less than the TVMAs during the first week of the wildfire, were competitive with the TVMAs in the subsequent weeks, and by week seven, some of the individual AEMs had coverage probabilities that were slightly higher than the TVMAs. The AEMs contained the final fire size between an average of 75 and 78 percent of the time over the first 60 days depending on the individual model (Figure 4.6). The EOFMs had very low coverage probabilities during the early weeks that increased over time and at no point had coverage probabilities that exceeded the TVMAs. The EOFMs contained the final fire size between an average of 0.4 and 0.46 percent of the time over the first 60 days depending on the individual model (Figure 4.6).

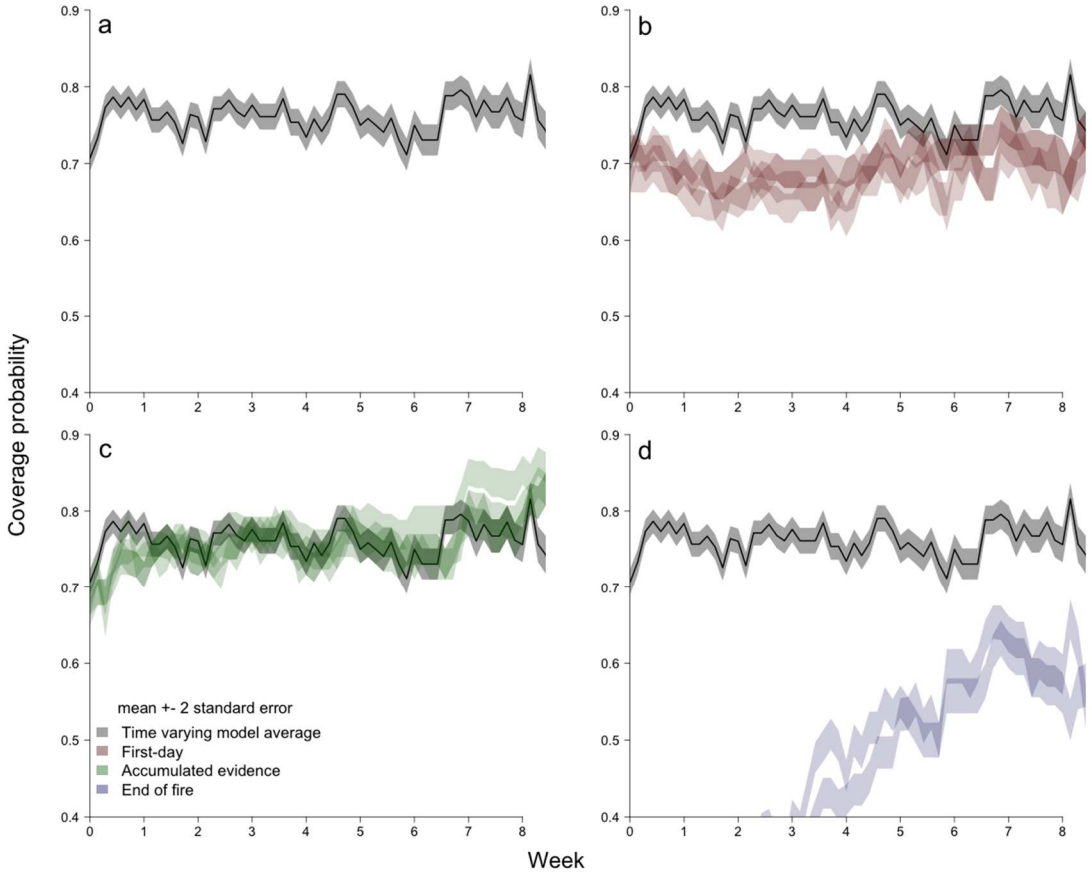


Figure 4.6. The probabilities that the central 80 percent of the posteriors contained the reported final fire size, the coverage probabilities, over time for the time varying model average (a), first-day models (b), accumulated evidence models (c), and end-of-fire models (d). Time is shown on the x-axis and is relativized for each fire so that values near zero represent the coverage probability of forecasts near the ignition date. In general, the TVMAs contained the reported final fire size about 75 percent of the time, which were rarely exceeded by the individual model constituents. Some of the AEMs had higher coverage probabilities in the late weeks of the fire.

The conditional probabilities that a wildfire will become a VLF – more than 4,047 ha – given it eventually did, increases over time regardless of the model. On the first day of the wildfire, these conditional exceedance probabilities are about 0.2 under the TVMAs, FDMs, and AEMs. Although the EOFMs cannot be used on the first-day of the fire, since they require growth information, the first predictions from these models produce conditional probabilities that are near-zero. Over time these probabilities increase at similar rates to the FDMs and AEMs, so

that the conditional probability that a wildfire becomes a VLF given it eventually did are nearly the same regardless of the choice of individual model. The conditional probability that a wildfire will become a VLF given it never did differently varied over time depending on the choice of model class. Specifically, the TVMA model conditional probabilities decreased over time, so that predictions became increasingly likely that a wildfire would not be a VLF. This same trend was not observed in the FDMs, AEMs, and EOFMs. For the first 60 days of the wildfire, the individual FDMs predicted that a VLF would occur with a probability of 19.1 to 19.9 percent when they did not occur; the individual AEMs predicted that a VLF would occur with a probability of 6.7 to 7.3 percent when they did not occur; and the individual EOFMs predicted that a VLF would occur with near-zero (<1 percent) probabilities when they did not occur. In contrast to the TVMAs, the conditional probability of a VLF given one did not occur did not strongly vary over time in the FDMs, AEMs, and EOFMs.

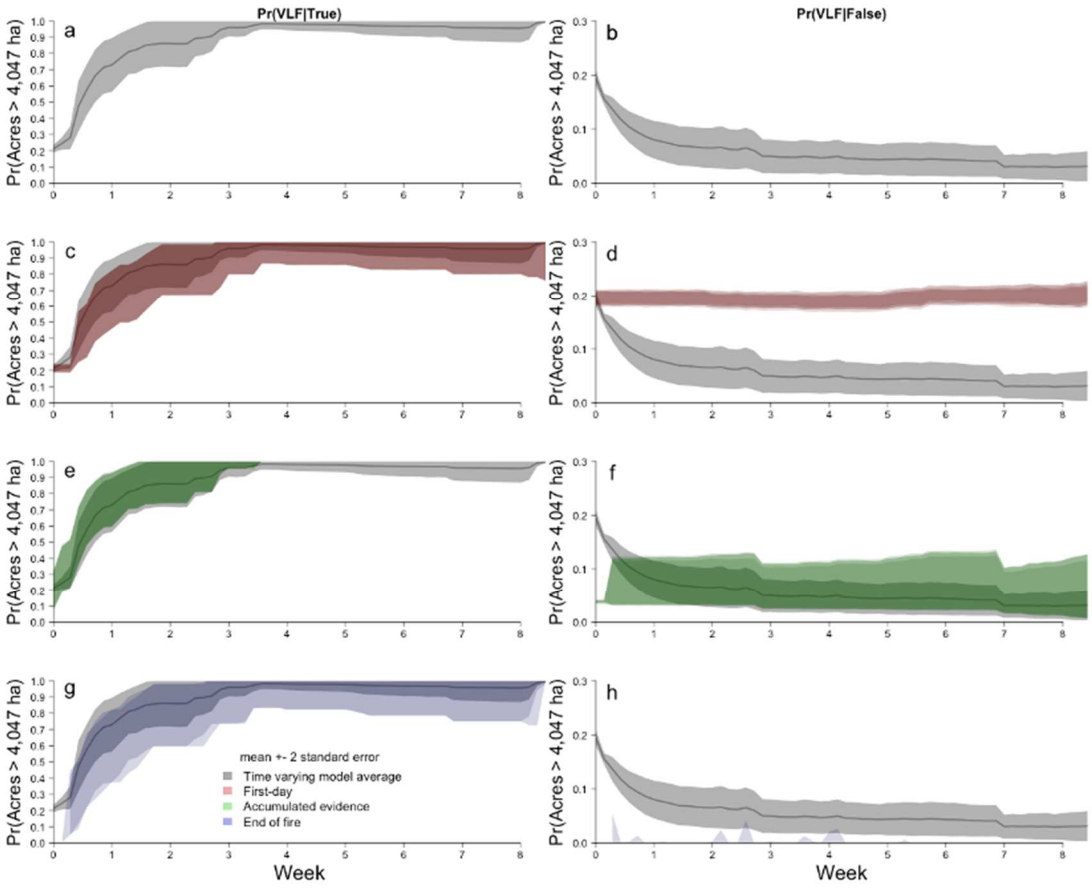


Figure 4.7. The average conditional probabilities that a very-large fire would occur over time when they eventually did occur (a,c,e,g), and when they never did occur (b,d,f,h). Panels (a) and (b) represent the time varying model average, panels (c) and (d) represent the first-day models, panels (e) and (f) represent the accumulated evidence models, and panels (g) and (h) represent the end-of-fire models. The typical conditional probabilities that a very-large fire would occur when they did followed were approximately similar temporal trends across the TVMAs and individual model components. However, the temporal trends varied across models when considering the conditional probability that a very-large fire would occur when they did not occur.

The interquartile range decreased over time in the TVMAs. The interquartile range also tended to decrease in the FDMs, although it approached a limit larger than that of the TVMAs. In other words, the posterior of TVMAs was less variable than the FDMs. The interquartile range of the AEMs was approximately constant over time, and like the FDMs, tended to be wider than the TVMAs late in the wildfire's lifetime. Like the TVMAs and FDMs, the interquartile range of the EOFMs decreased over time, but was consistently less than the TVMAs.

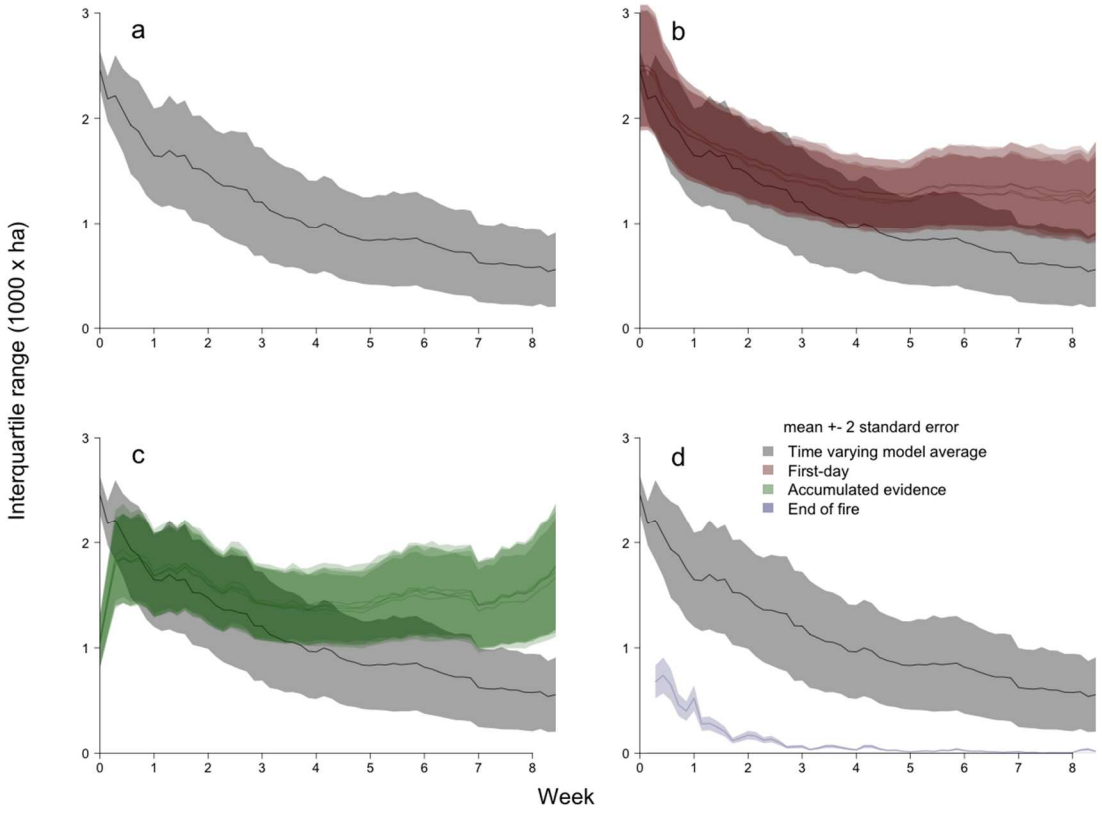


Figure 4.8. Interquartile range for the TVMA and individual constituent models over time for the time varying model average (a), first-day models (b), accumulated evidence models (c), and end-of-fire models (d).

4.4 DISCUSSION

4.4.1 *Benefits of TVMA*

When predicting final wildfire size, there is good reason to believe that different predictors should be utilized at different times. For instance, data is fairly limited in the first-days of a wildfire and may not be representative of the conditions later on in the wildfire's lifetime, but also represents the best available information at the time. Later on, higher quality information becomes available which can improve predictive ability. For this reason, different models should be utilized at different points at different times of a wildfire's lifetime. Additionally, even at one

fixed point of time there is uncertainty as to what model form should be used to generate predictions. The metamodel I described accounts of the time-varying and base level model uncertainties and I found that the metamodel generally improved predictive ability compared to any single model.

Specifically, the TVMA performed well across time and a range of metrics although the individual model components could outperform them in some specific contexts. For instance, although some of the individual AEMs had higher coverage probabilities than the TVMAs (Figure 4.6c), the AEMs also tended to overestimate wildfire sizes (Figure 4.5c), failed to correct VLF probabilities in wildfires that were increasingly unlikely to become VLFs (Figure 4.7f), and were less precise than TVMAs (Figure 4.8c). Similarly, the EOFMs were more accurate (Figure 5d) and precise (Figure 4.8d) late in the wildfire's lifetime but provided extremely poor size estimates the rest of the time (Figure 4.5d, Figure 4.6d). Hence, although there exist instances where some of the individual constituent models are superior during a particular moment in a wildfire's lifetime according to a specific performance measure, the TVMAs provide predictions that generally outperform any of the individual models overall, which is a result consistent with others (Raftery and Zheng 2003). One major reason this is a plausible explanation is that as the wildfire grows, an increasing lower threshold is applied to the posterior of the FDMs and AEMs while the upper tails are left unconstrained. When the wildfire growth eventually stops, the FDMs and AEMs continue to predict that more growth will occur, which is an artifact of the lower censored distribution (Alvarado et al. 1998). This explains why predictions from the individual FDM and AEMs are closer to the final fire size shortly following the ignition than later on in the wildfire's lifetime (Figure 4.5b, Figure 4.5c). In contrast, the TVMAs are able to

constrain the upper tails via the EOFMs, and thus provide predictions of final fire size that are fairly robust over time. Hence, the TVMA metamodel described here provides a means of integrating information from multiple models to produce predictions that are simultaneously accurate – compared to the individual model components - and realistically describe the level of uncertainty.

4.4.2 *Limitations*

Although the TVMA provided predictions that were generally superior compared to the individual constituent models, the exceptions to this result suggests that decision makers should still have individual predictions made available in parallel to the TVMA predictions. Some users may have evidence that one model class is particularly relevant. For instance, in situations where a wildfire is known to be nearly extinguished, end-users may prefer to consult the EOFMs than the FDMs or AEMs. An alternative metamodel could then modify the metamodel to advance through the model classes using relevant covariates. For instance, the time-varying model weights could instead be based on containment percentages, which are readily available for most wildfires in the United States (National Interagency Fire Center 2020). Slightly modifications to the TVMA metamodel can be used to recalibrate the coverage probabilities (Raftery 2016), although the difference between the expected coverage probabilities (80 percent) and actual (76 percent) are small enough that this may not be necessary.

In addition to modifying the metamodel, the Bayesian model averaging techniques explored here easily allow for the substitution of other model to predict final fire size. Other important factors that may be relevant to wildfire growth such as vegetation characteristics (Bradley et al. 2016), fire suppression policy (Brotons et al. 2013), or time since last fire (Youcom et al. 2019, Stevens-Rumann et al. 2016), could be used to improve the predictive

ability of these individual models. Other model structures, such as decision trees (Coffield et al. 2018), could be used instead, and in the case of GLMs, other combinations of link functions and probability distributions could be explored (Müller and Stadtmüller 2005). Fire size estimates produced from complex wildfire simulators could even be included within this metamodel (Finney et al. 2011). Although the model sets used for the FDMs and AEMs were selected as to incorporate the uncertainty as to what performance measures are relevant to consumers of the model predictions, other methods of generating models could be used. The model sets could have been limited to vertices of the barycentric plots (Figure 1) so that only the models with the lower errors, highest likelihood, and highest recall were considered (Podschwit et al. 2020a). Akaike weights could instead have used to identify the model set with a specified probability of containing the best-approximating model of wildfire size for the FDMs and AEMs (Symonds and Moussalli 2011). These changes to ensembles could potentially improve model performance and/or reduce the computational burden associated with fitting the metamodel parameters. The latter could also be reduced by using alternative Markov Chain Monte Carlo methods such as Hamiltonian MCMC (Podschwit et al. 2018b).

The presence of measurement uncertainties is an unavoidable limitation of working with burned area data and should be communicated to end-users. Missing and erroneous data is a well-known and common characteristic of burned area time series (Podschwit et al. 2018a) including the data used for validation (Podschwit et al. 2020b). Although robust methods exist for correcting these issues, this can be time consuming and computationally expensive (Podschwit et al. 2018a). For that reason, the methods adopted in this study were relatively simple (Denis et al. 2020). Still, measurement errors observed in burned area time series tend to

be low and were deemed tolerable for the purposes of this study, but future work could employ more sophisticated data cleaning methods to verify that this assumption.

4.4.3 *Recommendations for use*

As mentioned in the introduction, one obstacle to producing useful wildfire forecasts is the uncertainty associated with how the predictions will be used by end-users. Casual consumers that are not strongly affected by variation about the estimates could use the median posterior prediction and be fairly close to the final wildfire size on average. Worst case predictions on day 4 support areal estimates within 10.5 percent of the reported final fire size on average. For comparison, radial errors of $\pm$35% in wildfire spread models are considered good, which would correspond to an areal error of $\pm$82%; and radial errors of $\pm$10% would be considered very good, which would correspond to areal errors of $\pm$21% (Cruz and Alexander 2013).

Still, many users are strongly impacted by errors and may desire information regarding the credibility in the reported fire size estimates. In such cases, the interval predictions can provide this context with an estimate of the final size and a range of values that are likely to contain the reported final size. The 80 percent predictive intervals were only slightly miscalibrated, containing the reported final fire size an average of 76 percent of the time. As mentioned in the limitations subsection, the intervals could be corrected, but the difference between the expected coverage probability and the value estimated from the validation set is fairly small. The intervals have the added benefit of shrinking over time so that they increasingly limit the range of values that the fire is anticipated to grow to over time. In contrast, the first-day models are static over time and may report an intervals that are relatively wide (e.g. 80% probability that a wildfire is between 1 ha and 10^5 ha) compared to the TVMA intervals later on the wildfire's lifetime.

Lastly, some users rely on an ability to estimate the chance that specific size thresholds are likely to be exceeded by a wildfire. In this case, the TVMA metamodel identifies wildfires likely to become VLFs about as well the constituent models. However, the TVMA does better than the individual metamodel at identifying wildfires that are not likely to grow to be VLFs. Hence, the use of individual models may be adequate or not depending on the framing of predictive problem. That is, if filtering out non-VLFs is important then the TVMA metamodel should be preferred. However, if the identification of VLFs is the goal of the end-user, or the exceedance of any other arbitrary size threshold, then the TVMA performs about as well as the individual models.

That the performance of constituent models is often comparable to complex wildfire simulators (see Chapter 3) and that the TVMA approach further improves performance, suggests that decision makers can make better decisions utilizing this and similar approaches. Further refinement of these models may allow semi-automated decision-making in firefighting, where resources are allocated to reduce overall costs. Management actions could be included as covariates and the costs estimated from the location of the fire, allowing then comparisons between the damages of the resulting fire, which is also a function of management actions with associated expenditures. Although this would require clearly defining which costs are relevant (timber losses, property damages, human lives, ecosystem services, etc.), this is possible to do and would have potential to increase the efficiency of firefighting.

4.1 CONCLUSIONS

Model averaging can provide fire size forecasts that are generally superior to the individual constituent models. Model averaging can be used to combine wildfire size forecasts that utilize varying levels of information. Initial forecasts based on information available on the first-day can

be produced first, which are later improved forecasts that utilize growth information that is only available later on in the wildfire's lifetime. A number of modifications could be adopted to improve the predictive ability of the metamodel - such as using covariates in the metamodel and other model ensembles - and future work should explore how these changes alter the predictive ability. Still, the metamodel described in this manuscript already show promising predictive ability relative to the individual constituent models and may serve as a template from which these modifications can build.

4.1 REFERENCES

Alvarado E., Sandberg D.V. & Pickford S.G. (1998) Modeling large forest fires as extreme events. *Northwest Science.* 72:66–75

Bailey, R.G. (2016). Bailey's ecoregions and subregions of the United States, Puerto Rico, and the US Virgin Islands. Forest Service Research Data Archive: Fort Collins, CO, USA, 2016. Available online: https://doi.org/10.2737/RDS-2016-0003 (accessed on 14 December 2018).

Bradley, B.A., Curtis, C.A., & Chambers, J.C. (2016). Bromus response to climate and projected changes with climate change. In Exotic Brome-Grasses in Arid and Semiarid Ecosystems of the Western US (pp. 257-274). Springer International Publishing.

Brooks, S.P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics.* 7(4), 434-455.

Brotons, L., Aquilué, N., De Cáceres, M., Fortin, M.J., & Fall, A. (2013). How fire history, fire suppression practices and climate change affect wildfire regimes in Mediterranean landscapes. *PLOS one*. 8(5), e62392.

Coffield, S.R., Graff, C.A., Chen, Y., Smyth, P., Foufoula-Georgiou, E., & Randerson, J.T. (2019). Machine learning to predict final fire size at the time of ignition. *International Journal of Wildland Fire.* 28(11), 861–873.

Cruz, M.G., & Alexander, M.E. (2013). Uncertainty associated with model predictions of surface and crown fire rates of spread. *Environmental Modelling & Software*. 47, 16-28.

Cui, W., & Perera, A.H. (2008). What do we know about forest fire size distribution, and why is this knowledge useful for forest management?. *International Journal of Wildland Fire.* 17(2), 234-244.

Cullen, A.C., Axe, T., & Podschwit, H. (2020). High-severity wildfire potential–associating meteorology, climate, resource demand and wildfire activity with preparedness levels. *International Journal of Wildland Fire.* 29(12)

Denis, L. A.S., Mietkiewicz, N.P., Short, K.C., Buckland, M., & Balch, J.K. (2020). All-hazards dataset mined from the US National Incident Management System 1999–2014. *Scientific Data.* 7(1), 1-18.

Finney, M.A., Grenfell, I.C., McHugh, C.W., Seli, R.C., Trethewey, D., Stratton, R.D., & Brittain, S. (2011). A method for ensemble wildland fire simulation. *Environmental Modeling & Assessment*, 16(2), 153-167.

Fragoso, T.M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review.* 86(1), 1-28.

Heim Jr, R.R. (2002). A review of twentieth-century drought indices used in the United States. *Bulletin of the American Meteorological Society.* 83(8), 1149-1166.

Intini, P., Ronchi, E., Gwynne, S., & Pel, A. (2019). Traffic modeling for wildland–urban interface fire evacuation. *Journal of Transportation Engineering.* 145(3), 04019002.

Kolden, C.A., & Weisberg, P.J. (2007). Assessing accuracy of manually-mapped wildfire perimeters in topographically dissected areas. *Fire Ecology*. 3(1), 22-31.

Meyn, A., White, P.S., Buhk, C., & Jentsch, A. (2007). Environmental drivers of large, infrequent wildfires: the emerging conceptual model. *Progress in Physical Geography.* 31(3), 287-312.

Miller, J.D., Safford, H.D., Crimmins, M., & Thode, A.E. (2009). Quantitative evidence for increasing forest fire severity in the Sierra Nevada and southern Cascade Mountains, California and Nevada, USA. *Ecosystems.* 12(1), 16-32.

Moeltner, K., Kim, M.K., Zhu, E., & Yang, W. (2013). Wildfire smoke and health impacts: A closer look at fire attributes and their marginal effects. *Journal of Environmental Economics and Management.* 66(3), 476-496.

Morgan, M.G., Henrion, M., & Small, M. (1990). Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press.

Müller, H.G., & Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics.* 33(2), 774-805.

National Interagency Fire Center. (1 October 2020) *ICS-209 Program (NIMS).* 2020 ICS-209 User Guide. https://www.predictiveservices.nifc.gov/intelligence/ICS-209_User_Guide_4.0_2020.pdf

Raftery, A.E. (2016). Use and communication of probabilistic forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal.* 9(6), 397-410.

Raftery, A.E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*. 98(464), 931-938.

R Core Team (2019) R: A language and environment for statistical computing. (Vienna, Austria) Available at http://www.Rproject.org [Verified 1 September 2020].

Stevens-Rumann, C.S., Prichard, S.J., Strand, E.K., & Morgan, P. (2016). Prior wildfires influence burn severity of subsequent large fires. *Canadian Journal of Forest Research.* 46(11), 1375-1385.

Podschwit, H., & Cullen, A. (2020). Patterns and trends in simultaneous wildfire activity in the United States from 1984 to 2015. *International Journal of Wildland Fire*. 29(12), 1057-1071.

Podschwit, H., Guttorp, P., Larkin, N.K. Estimating wildfire growth from noisy and incomplete incident data using a state space model. *Environmental and Ecological Statistics.* 25, 325–340 (2018a). https://doi.org/10.1007/s10651-018-0407-5

Podschwit H., Larkin N.K., Steel E.A., Cullen A., Alvarado E. (2018b) Multimodel forecasts of very-large-fire occurences during the end of the 21st century. *Climate.* 6, 100. doi:10.3390/CLI6040100

Podschwit, H., Larkin N.K., and Steel E.A. Wildfire growth rates may improve predictive performance of statistical forecasts of final fire size. *Environmental Modeling & Assessment* (2020a) *In review.*

Podschwit, H., Potter, B., and Larkin N.K. A protocol for collecting burned area time series cross-check data with application. *Fire* (2020b). *In preparation.*

Riley, S.J., DeGloria, S.D., & Elliot, R. (1999). Index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, *5*(1-4), 23-27.

Sanderson, E.W., Jaiteh, M., Levy, M.A., Redford, K.H., Wannebo, A.V., & Woolmer, G. (2002). The human footprint and the last of the wild: the human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience,* 52(10), 891-904.

Symonds M.R., Moussalli A. (2011) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology.* 65, 13–21. doi:10.1007/S00265-010-1037-6

Taylor, J.G., Gillette, S.C., Hodgson, R.W., & Downing, J.L. (2005). Communicating with wildland interface communities during wildfire (No. 2005-1061). US Geological Survey.

Thompson, M., Calkin, D., Scott, J.H., & Hand, M. (2017). Uncertainty and probability in wildfire management decision support: an example from the United States. Natural hazard uncertainty assessment: modeling and decision support. *Geophysical Monograph*, 223, 31-41.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*. 63, 24-31.

Weiss, A., & Hays, C.J. (2005). Calculating daily mean air temperatures by different methods: implications from a non-linear algorithm. *Agricultural and Forest Meteorology*. 128(1-2), 57-65.

Yocom, L.L., Jenness, J., Fulé, P.Z., & Thode, A.E. (2019). Previous fires and roads limit wildfire growth in Arizona and New Mexico, USA. *Forest Ecology and Management*. 449, 117440.

Chapter 5. CONCLUSIONS

I have found that model uncertainty is an aspect of wildfire prediction that cannot simply be ignored. In Chapter 2 for instance, I found that in some cases very-large fire activity could be predicted to either increase or decrease depending on which combination of climate model and predictive model selected, and that only by integrating the results across these combinations could I can be confident that they were robust to uncertainties in both the predictive models and climate models. Similarly, in Chapter 3 - where instead of the presence/absence of very-large wildfires in the future, I examined the effects of model uncertainty in real-time statistical forecasts wildfire size - I found that substituting utility functions could sometimes result in new predictive models that could change predictions of wildfire size by multiplicative factors of more than 100. On the other hand, this sensitivity was not always observed. In Chapter 2, when predictive models were based on covariates that did not vary strongly across climate models, the effects of model uncertainty were relatively small and the eschewing the integration of results across predictive models and climate models would not have drastically altered the conclusions. Moreover, in Chapter 3, the first-day model predictions - which were based on models that use meteorological, topographic, and anthropogenic covariates - were robust and produced similar predictions regardless of which utility functions was considered. Hence, without investigating the effects of model uncertainty on predictions the credibility of wildfire predictions may be in doubt. If such analyses are not completed, one cannot know whether our predictions will be drastically change if new information is considered or not. Although in some cases the effects of model uncertainty were small, in other cases the effects of model uncertainty on prediction were large enough that ignoring this uncertainty would be careless if our goal is to inform important and costly wildfire-related decisions. For example, in Chapter 2, some of the combinations of

predictive models and climate models yielded predictions of decreases of wildfire activity even though in aggregate they suggested increases. A decisionmaker basing decisions on the information in the former may be too risk tolerant, when risk aversion was warranted. One example of a negative consequence arising from unjustified risk tolerance would be training too few firefighters and incident commanders to manage the future increases in very large fire activity. In contrast to overconfidence, underconfidence is another insidious problem in prediction that can be guarded against by assessing model uncertainty. To elaborate, consider that the predictions from the first-day models were, at times, very wide, offering predictions of fire size that ranged by tens of thousands of acres. If other models were not considered, analysts would be stuck with this wide intervals, but by looking at the accumulated evidence models and end-of-fire models, these intervals could be shrunk providing actionable information to decisionmakers.

Given the importance of assessing levels of model uncertainty in wildfire prediction, I also described methods for incorporating this uncertainty ranging from simply reporting the predictions from each model independently (Chapter 3) to integrating the predictions into a single distribution or prediction using metamodeling (Chapter 2 and Chapter 4). Although presenting the models independently is a simple approach that can be used by most decision makers, it also results in difficult to interpret results. In Chapter 3, enumerating the models and their associated predictive performance required large tables of information and complex figures, which made difficult concisely describing what covariates were important where and how predictive performance varied across the models. The challenges with interpretability associated with this approach may be insurmountable in some applications. In Chapter 2 for instance, consulting all 1300 model combinations would be impractical, and some form of prediction

aggregation would almost certainly be needed. Hence, in addition to the observed performance benefits, the use of Bayesian model averaging is perceived as an improvement in that the results are easily interpreted. Instead of trying to compare multiple models individually, the set of models was integrated into a single prediction or distribution that could be understood by lay-audiences with little explanation. This technique would then have the advantage of allowing decision-makers to consume predictions of wildfire size that incorporate model uncertainty in ways that are already similar to the methods that they use to consume predictions from a single model. Instead of decisionmakers attempting to use intuitions - intuitions that can be easily misguided - to make sense of the results from multiple contradictory models, the metamodel approach does this accounting for them so it is as if the decisionmaker is using a single model. The principle disadvantage of this approach is that it requires technical skills to implement that are usually outside the training of the decision-makers it is intended to assist but is a problem that is surmountable with outside consultation.

It should not be expected that the effects of model uncertainty on wildfire prediction are limited to statistical models only. These problems certainly extend into complex fire and smoke modeling, which utilize far more parameters and are often explicitly based on physical relationships rather than statistical. The similar performance of FSPro compared to the simple statistical models developed in Chapter 3 demonstrates some of the challenges of using single model approaches. That is, if highly complex models produce similar predictions to simple models, then model users may question what all the additional complexity has bought them. Even if it is accepted that the additional complexity of these models is justified, the results of Chapter 3 suggest that the optimal model is likely to be highly sensitive to the method used to gauge performance. Even if the models already used in the real-world in firefighting operations

and smoke forecasting are optimal under a specific criterion, they will likely change when another criterion is considered. Similarly, complex models may be subject to time varying levels of information availability which may influence model selection. Analogous time sensitive patterns in covariate use that were observed in Chapter 3 may extend to complex models. One could imagine a situation where one model provides superior performance, but also requires fire perimeter information that is unavailable in the early days of a fire. A competing model may provide somewhat worse performance in the early days of the fire, but also provides the best guess given the information available. on when little time has elapse since ignition, but these same models may when a wildfire just ignites some models may provide superior performance during the early. The problems described here could be addressed using similar methodologies as those described in this dissertation. However, unlike the models described here, many of these smoke and wildfire spread models produce maps that would require modifications to the metamodels described here to implement including defining the relevant utility function (e.g. likelihood) and identifying new model ensembles (e.g. FSPro and Farsite). Still, the final outcome would still be a weighted average of predictions from multiple models with weights that optimize a well-defined utility function. Moreover, the detailed predictions and complex models will require much more computational resources than those used in these analyses, which may necessitate the use of faster computational methods. Data availability could also be a problem in these future analyses as growth progression data are notoriously rare. However, as overconfidence is particularly dangerous when the costs are high, and the decisions informed by these models are very important to protecting lives and property, the future exploration of model uncertainty in these contexts could be immensely valuable to the real-world decisionmakers that

use these complex models. The analysis of model uncertainty permits decisionmakers to make more informed choices than is allowed when model uncertainty is ignored.

Another aspect that was little explored in this dissertation, but is deserving of further study, is how this quantitative information is interpreted by decision makers. In this dissertation, I explored to general methods of presenting model uncertainty 1.) presenting model information independently and 2.) integrating results using a metamodel. It was assumed that (2) would be easier to interpret, however verifying this is important as the computational overhead associated with the latter is of little benefit if in the end the information is misinterpreted by the decisionmakers it is intended to assist. Moreover, there may exist contexts in which decision makers may prefer the former method. In Chapter 4 for instance, it was acknowledged that the end-of-fire models will outcompete the model averaged results at some point in the wildfire's lifetime, meaning that better predictions could be produced if people knew exactly when to listen to the end-of-fire models. Hence, future work should then determine if the benefits of interpretability metamodeling are valid and in what contexts they are most applicable. Future research should also identify effective ways of presenting this quantitative information to minimize the chances of misinterpretation. This will involve investigating the psychology of decision makers and thus require a set of methods that are wholly distinct from those used in these statistical analyses.

Due to similar predictive performance across models, uncertainty regarding the choice of utility function, or time-varying reliability of data inputs, one often cannot confidently identify a single model that should be consulted in all cases. I however provide tools for identifying the levels of model uncertainty and have for the first time in fire modeling history applied metamodel frameworks to integrate this uncertainty so that decisions can be made that avoid the

overconfidence that is so dangerous when attempting to forecast, an oftentimes a difficult to predict, future.

VITA

Harry Podschwit took an interest in fire early in his academic career, assisting with prescribed fires at the College of DuPage after graduating high school. Harry also developed a curiosity for how mathematics could be applied to problems in the natural sciences and transferred the Applied and Computational Mathematical Sciences program at the University of Washington to complete his undergraduate, specializing in mathematical biology. Following his undergraduate, Harry enrolled in the Quantitative Ecology and Resource management program at the University of Washington to complete his M.S. under Peter Guttorp, researching how statistical analysis could be used to improve wildfire growth data quality and predict wildfire behaviors.