Part II: Lessons Learned from Predicting Wildfire Occurrence for CONUS Using Deep Learning and Fire Weather Variables

BETHANY L. EARNEST,[a,b,e] AMY MCGOVERN,[a,b,c,d] CHRISTOPHER KARSTENS,[e,d] AND ISRAEL JIRAK[e]

[a] School of Computer Science, University of Oklahoma, Norman, Oklahoma
[b] Cooperative Institute for Severe and High-Impact Weather Research and Operations, Norman, Oklahoma
[c] NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, Norman, Oklahoma
[d] School of Meteorology, University of Oklahoma, Norman, Oklahoma
[e] Storm Prediction Center, Norman, Oklahoma

ABSTRACT: This paper illustrates the lessons learned as we applied the U-Net3+ deep learning model to the task of building an operational model for predicting wildfire occurrence for the contiguous United States (CONUS) in the 1–10-day range. Through the lens of model performance, we explore the reasons for performance improvements made possible by the model. Lessons include the importance of labeling, the impact of information loss in input variables, and the role of operational considerations in the modeling process. This work offers lessons learned for other interdisciplinary researchers working at the intersection of deep learning and fire occurrence prediction with an eye toward operationalization.

KEYWORDS: Wildfires; Deep learning; Neural networks; Machine learning; Artificial intelligence; Data science

1. Introduction

In Part I of this work, Earnest et al. (2024), we developed and analyzed a U-Net3+ deep learning model for the task of fire occurrence prediction in CONUS in the 1–10-day range. We developed two models, the "All Fires" model which predicted the probability of fire of any size and any cause and the "Large Lightning" model which only predicted the probability of lightning-caused fires greater than or equal to 1000 acres in final fire size. We demonstrated that the All Fires model performed better than the Large Lightning model in general, on large lightning fires in specific, and on the largest of large lightning fires for the test dataset. In Part II of this work, we explore the reasons why the All Fires model outperformed the Large Lightning model and offer lessons learned that may be useful to other interdisciplinary researchers, specifically for those developing machine learning applications for wildfire prediction.

2. Lessons learned

In the following section, we discuss decisions made during the modeling process that either improved or limited model performance. In section 2a, where we explore model performance improvements, we discuss why the models performed the way they did even in the absence of lighting inputs. In section 2b, where we explore choices that limited the model in some way, we discuss how variable selection and codification can limit

model performance or remove the model's ability to offer certain insights all together.

a. Where is the lightning?

One critical point worth drawing out is that neither the All Fires model nor the Large Lightning model had lightning observations as an input. This seems to be a condemning omission for how can lightning fires be predicted without the model having knowledge of lightning? Exploring this question will help illustrate why performance differed between the All Fires model and the Large Lightning model.

1) LESSON 1: HIGHER LABEL DENSITY LEADS TO BETTER PERFORMANCE

The first lesson we took away from Part I is that more labels lead to better performance. This is a different lesson than the commonly known "more data leads to more performance" because both models were given the same amount of data (same domain in time, space, and input variables). Where the models differed is in the number of instances of fire occurrence depicted in the label images used to train the models and to measure model performance. Since the All Fires model included fires regardless of size or cause, more instances of fire occurrence were available for the label images offered to the model as depicted in Table 1. This provided the All Fires model with more opportunities to identify meaningful patterns between the inputs and the labels which may have allowed it to be more performative. The All Fires model was able to issue fire probability predictions both where large lightning fires were discovered and where other causes and sizes of fires were discovered resulting in better extrapolation when applied to the test dataset.

Corresponding author: Bethany L. Earnest, bethany.earnest@gmail.com

TABLE 1. Fire occurrence label counts by season for the All Fires model and Large Lightning model.

Cross-validation rotation 1	Season	All fires label count	Large lightning fires label count
Train (2000–14)	Fall	216 440	269
	Spring	412 083	268
	Summer	402 630	3024
	Winter	202 069	9
Validation (2015–17)	Fall	49 953	51
	Spring	75 108	46
	Summer	76 064	656
	Winter	38 594	0
Test (2018–20)	Fall	38 776	41
	Spring	75 574	41
	Summer	76 821	489
	Winter	24 974	2

Examples of this relationship can be seen in Fig. 1, where a gap in performance can be observed between the All Fires model and all other models attempted, each of which sought to constrain the cause or size of the fire occurrence instances included in the label images in some way. We validate in the following section that this increase in performance does not come at the expense of maintaining high performance on specific cases of interest, e.g., large lightning fires.

2) LESSON 2: START WITH THE GENERAL CASE AND THEN GO SPECIFIC, NOT THE OTHER WAY AROUND

We framed the task of predicting large lightning fires into the following three components:

Component 1: A cloud-to-ground (CG) lightning flash must meet the criteria necessary to start a fire: it must contact the ground, it must connect with something combustible, and it must have high enough amperage and low enough voltage to convert electrical charge to combustion (Pyne 2001).

Component 2: Fire mitigating weather (such as frequent, heavy rains) must not be present as it hinders the ignition and spread of fire.

Component 3: CG lightning must occur in an area that is already primed for fire.

We can see from the framing presented that the scope of each component builds on the scope of the previous component. Component 1 deals only with the ignition source. Component 2 deals with the means of mitigation for the phenomenon of interest. Component 3 deals with the larger environmental factors which can contribute to making an area suitable for the phenomenon of interest. It is tempting to focus first on those components that apply only to large lightning fire, components 1 and 2, so as not to dilute the signal received by the model. This is a lesson carried over from numerical weather prediction practices common in the weather and fire weather spaces. It does not necessarily apply when using a deep learning model.

Instead, as we demonstrate in Part I (Earnest et al. 2024), by taking the opposite approach, and starting with the most general component, component 3, we were able to build a performative model (the All Fires model), both in general and on the specific case of interest, large lightning fires. This performance stems from the All Fires model's ability to
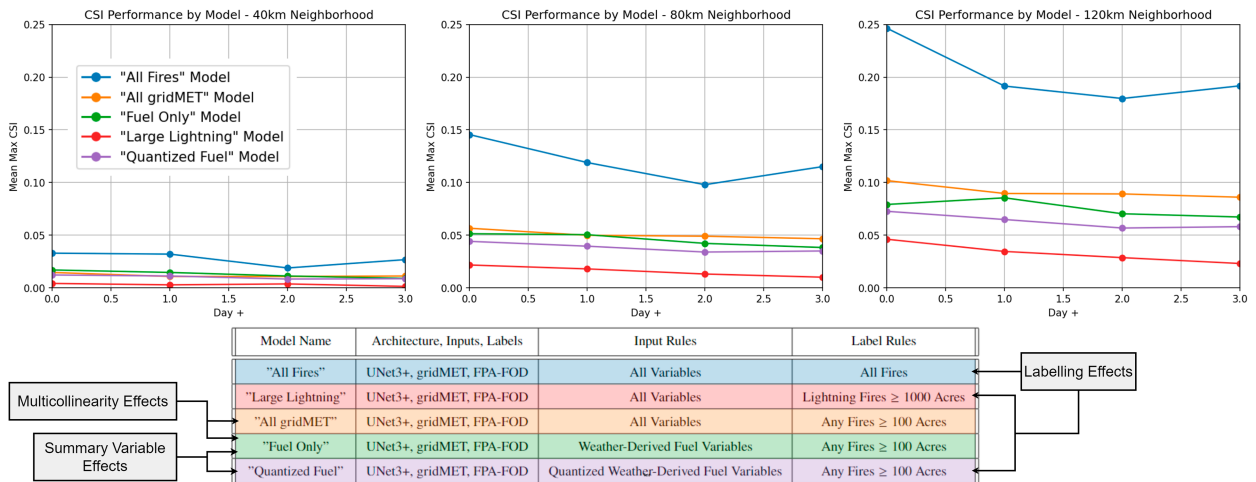

FIG. 1. Max CSI by model.

predict where fires, any fires, are likely to be discovered in CONUS, or, said differently, which areas of CONUS are primed for fire. In contrast, the Large Lightning model predicts only where large lightning fires are likely to be discovered in CONUS, which provides a much narrower view into potential future fire behavior. Predicting where fires are possible, the goal of the All Fires model, as opposed to where large lightning fires are possible, the limitation of the Large Lightning model, is a key prospective shift necessary in building a model that generalizes well and has some measure of future proofing. After all, where large lightning fires have happened in the past does not dictate, exclusively, where large lightning fires, or indeed, all fires, will happen in the future.

Once a working understanding has been developed, an interdisciplinary researcher must keep in mind not only what methods have been used to address this challenge in the past but how those methods differ from deep learning so as to offer the most value from a deep learning implementation for the task at hand. In our work, that meant exploring a path that ran counter to previous, transferable modeling knowledge present in the weather and fire weather domain so as to offer the most performative deep learning model for the task of predicting wildfire occurrence.

b. There are inputs and then there are inputs

There are many ways to support the performance of a machine learning model. As discussed in the previous section, one way focuses on the impacts of labeling and another on how the modeler interprets the modeling space. Another way to support model performance is by selecting the best inputs for the predictions that the model will need to make. Part of this task deals with variable selection, a topic touched on in both this section and in the next section. Another piece of this task, focused on in this section, deals with how the variables, once selected, are codified (preprocessed or transformed) for use by the model.

One example of how variables can be codified for the model is the process of normalization. When using neural networks, it is important to normalize the model inputs because it supports the model training process by shifting all inputs into the same scale which helps stabilize gradient descent (the method by which neural networks learn) which can result in faster model convergence to an optimal solution. Said differently, the model gets as good as it is going to get as quickly as possible. This is a helpful trait for models destined for operationalization. For this work, all inputs were normalized to values between zero and one.

1) LESSON 3: SUMMARY VARIABLES HAVE THEIR PLACE, JUST NOT NECESSARILY IN A DEEP LEARNING MODEL

At the confluence of variable selection and variable codification is the question of summary variables. It is tempting in interdisciplinary work to seek out summary variables carefully created by the subject matter experts working in the space, based on years of experience and rules of thumb. This approach is not without merit and can speak to important themes valuable to the modeling task, such as what variables forecasters consider to be important. However, when machine learning is not the currently used method, it is valuable to keep in mind not only what has worked for forecasters but also what will work for the machine learning model. In the case of summary variables, such as climatology variables and probabilities derived from climatology, the data components that human forecasters rely on do not necessarily provide the machine learning model with the complexity of data that it needs to produce its highest quality prediction.

One early approach we tried in applying machine learning to the task of fire occurrence prediction, informed by how subject matter experts use climatologies today, was to use quantized fuel variables, called the "Quantized Fuel" model in Fig. 1. We broke weather-derived fuel variables (energy release component, burning index, and dead fuel moisture 100 and 1000 h) down into quantiles (0, 10, 25, 50, 75, 90, and 100) and converted those quantiles into individual variables (e.g., one variable would be the 90th percentile burning index) and submitted those variables to the model as inputs, much as one would do when considering a climatology. The performance of the Quantized Fuel model, when compared to other models we developed as our work matured, was relatively low. Performance for all models is depicted in Fig. 1 in which model performance is measured using critical success index (CSI). CSI is a performance metric used to evaluate model predictions within the National Weather Service (Schaefer 1990).

Referring to Fig. 1 and comparing model performance for only models that used similar inputs and labeling strategies, the "Fuel Only" model outperformed the Quantized Fuel model. Both models used the same input variables from the same source [burning index, energy release component, and dead fuel moisture sourced from gridded surface meteorological (gridMET) dataset]; the difference between the two models was that the Quantized Fuel model relied on summary inputs (its input variables were converted to quantiles before being normalized and offered to the model) while the Fuel Only model relied on nonsummary inputs (its input variables were only normalized before being offered to the model).

One interpretation for this difference in performance is that the entire distribution, as described by the nonsummary input variables, carried more information than the values used to summarize the distribution, as described by the summary variables. The summary dependent model also took up more storage space and was more computationally intensive than the nonsummary dependent model as we added multiple summary variables (each with the same coverage in time and space but offering less information than the original variable) in an attempt to compensate for the information lost in summarization. This resulted in a large model which trained slowly and performed relatively poorly, characteristics which are less desirable for a model bound for operationalization.

The key take-away described by these results is that the use of summary variables as model inputs when the model is capable of handling the nonsummary variables results in a model starved for information. The appetite for summary variables stems largely from how the human mind works, famously
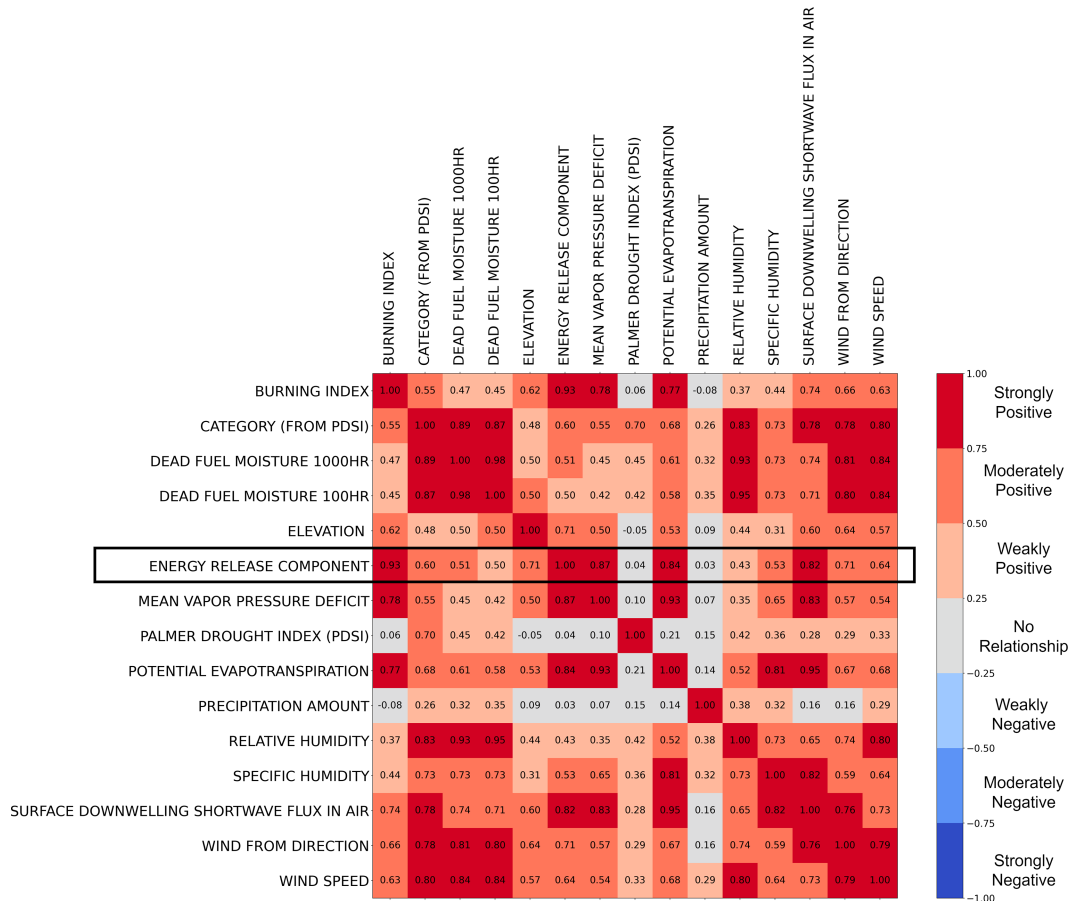
FIG. 2. Correlation plot of gridMET variables.

described by Tversky and Kahneman (1974) as heuristics, or the shortcuts the human mind takes to make decisions in the presence of uncertainty. A deep learning model needs no such shortcut and is either biased or constrained by being provided with one. An effective deep learning model can stand in for previously necessary shortcuts, allowing decision-makers to quickly connect the dots without having to sacrifice the amount of information considered. This trait makes deep learning a great candidate for decision support.

2) LESSON 4: MULTICOLLINEARITY AND VARIABLE CONTRIBUTION

Multicollinearity describes a situation wherein the inputs to a model are correlated with each other. While the presence of multicollinearity does not necessarily produce a less performative model, it does impact our ability to quantify the contributions of the individual input variables for the model. When multicollinearity is present, variable importance is difficult to measure because it is unknown from which correlated variable the effect on the model is produced. Possible solutions for such a situation are 1) to remove all of the highly correlated variables, retaining only the uncorrelated variables of interest, or 2) to apply dimensionality reduction [principal

component analysis (PCA) being an example of such] to reduce the amount of correlation in the dataset while retaining as much information as possible.

The datasets used for this work were recommended by subject matter experts because of their familiarity to fire weather forecasters with the goal of lowering barriers to adoption for the model. Between the input variables, sourced from grid-MET, there is a lot of correlation, as depicted in Fig. 2. While the rules defining which values of correlation correspond to different correlation strengths vary by situation and discipline, we have described our discretization in Fig. 2.

Our subject matter experts would like to receive information regarding variable importance in the context of the input variables used for the model and for those input variables to be variables familiar to them. For example, energy release component (ERC) is a popular variable used in fire weather forecasting which describes the amount of available energy [basic training unit (BTU)] per unit area (square foot) within the flaming front at the head of a fire (U.S. Department of Agriculture 2023). Fire weather forecasters would like to know how important ERC is to making an accurate fire weather forecast. With the All Fires model, perhaps, we have a model offering enough accuracy for the discussion to shift to variable importance.

Referring to Fig. 2, all but two input variables exhibit a positive correlation with ERC. It is important to understand that ERC is calculated using burning index and 1000-h fuel moisture, which is in turn calculated using temperature, relative humidity, and precipitation (National Interagency Fire Center 2023). This is called structural multicollinearity, wherein correlation between variables results from using one model term to create another model term, resulting in more variables for the model but not necessarily more information for the model.

Option 1, that of removing variables highly correlated with the variable of interest, is likely not at our disposal unless we are willing to lose the majority of our input variables. Option 2, that of applying dimensionality reduction to the input variable set, as in the use of PCA, would allow us to quantify variable importance but only in terms of new, linearly uncorrelated variables, called principal components. The resulting principal component variables would no longer be of a format recognized by the fire weather forecasters and would lack context save that offered by the model itself.

In summation, neither option 1 nor option 2 offers the ability to quantify variable importance within the context of recognizable variable inputs using the datasets recommended for this work as familiar to fire weather forecasters. As variable importance is a valuable tool for supporting model understanding, trust, and, ultimately, adoption, this creates a vulnerability for a model bound for operationalization.

3. Conclusions

From our work predicting fire occurrence using machine learning, we learned many lessons. In section 2a, we discussed two lessons learned that helped improve our model performance. First, even for rare events, increased label density can increase performance both on the general case of fire and on the specific case of large lightning fire. Second, by tackling the most general aspects of the problem space first, you can set the stage for a performative model that generalizes well both to previously unseen future data and to unexpected cases within the data (such as large lightning fires where you would not expect them to be). In section 2b, we discussed two lessons learned that limited our model in some way. First, we discussed the impact summary variables have on model performance when used as inputs. Though summary variables, such as climatologies and probabilities derived from climatologies, are an important tool in fire weather forecasting, when climatology equivalent variables were introduced as model inputs, model performance decreased. Second, we discussed the impact that multicollinearity among inputs has on the model's ability to provide desired insights which can be valuable to

model understanding and adoption. When multicollinearity is present, it is difficult to isolate variable contribution to model behavior and as variable contribution is a valuable insight for fire weather forecasters to have, the presence of multicollinearity represents a vulnerability for a model designed for operationalization.

Data availability statement. Data analyzed in this study were a reanalysis of existing data, which are openly available at locations cited in the reference section. Further documentation about data processing will be available prior to publication.

REFERENCES

Earnest, B., A. McGovern, C. Karstens, and I. Jirak, 2024: Part I: Improving wildfire occurrence prediction for CONUS using deep learning and fire weather variables. *Artif. Intell. Earth Syst.*, **3**, e230057, https://doi.org/10.1175/AIES-D-23-0057.1.

National Interagency Fire Center, 2023: Energy release component. Accessed 2 June 2023, https://gacc.nifc.gov/swcc/dc/azpdc/predictive-services/documents/fuels-fire-danger/ERC.pdf.

Pyne, S. J., 2001: *Fire: A Brief History*. University of Washington Press, 224 pp., http://www.jstor.org/stable/j.ctvcwnf8f.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Tversky, A., and D. Kahneman, 1974: Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124–1131, https://doi.org/10.1126/science.185.4157.1124.

U.S. Department of Agriculture, 2023: Energy release component (ERC) fact sheet. Accessed 2 June 2023, https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb5339121.pdf.