USFS-CSU Joint Venture Agreement Phase 1 (2017-2018): Developing a Gridded Model for Probabilistic Forecasting of Cloud-to-Ground Lightning Flashes over the Lower 48 States

Final Report

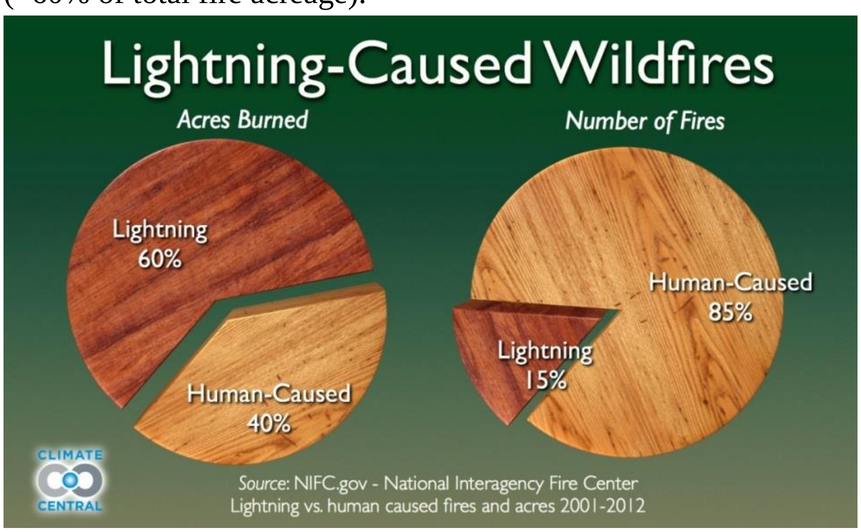
Ned Nikolov, Ph.D.
Phillip Bothwell, PhD.
John Snook, Ph.D.

February 2019

1. Introduction

The National Predictive Services (NPS) has asked the USFS Rocky Mountain Center for Fire-Weather Intelligence (RMC) managed by the WFM RD&A Unit of the USFS Rocky Mountain Research Station (RMRS) to assist with the development of a system of statistical models for predicting the ignition probability & growth potential of significant fires on a national grid based on weather forecasts. The development of this gridded system of predictive equations is planned to proceed in 3 stages. Phase 1 (Oct. 2017 – Dec. 2018) was aimed at developing and partial validation of a high-resolution spatial model for forecasting the probabilities of lightning flashes as a function of 3-D fields of atmospheric parameters stretching from the surface to the tropopause. Lightning is expected to be a strong predictor of fire activity in many parts of the Western US and Alaska. There are 3 main reasons for expending resources on the development of a state-of-the-art system of lightning-forecast equations:

a. Lightning has been an increasing cause for wildfires in recent years worldwide. Although historically lightning only accounts for about 15% of wildfire occurrences in the USA, lightning-ignited fires burn by far the most territory (~60% of total fire acreage).



b. There is a need for high-resolution operational lightning forecasts of superior skill out to 7-10 days based on long-term climatology that can be harnessed to predict ignition probabilities for naturally occurring wildfires.

c. A gridded Lightning Forecast Model is a prerequisite for the successful development of a new & improved "*7-Day Fire-Potential Outlook Products*" for Predictive Services.

The intended output from Phase 1 of this project is a set of logistic equations capable of predicting probabilities of one or more flashes and 10 or more flashes every 3 hours on a 20-km resolution national grid using 7-10 day numerical weather forecasts as input.

2. **Method**

This section describes the datasets and computational procedures utilized to develop and test the system of lightning forecast equations for the lower 48 states.

2.1. Gridded Data Sets

Table 1 lists the datasets utilized by Phase 1 of this project. Two 29-year long 3-h resolution climatological records of weather reanalysis fields and observed lightning flashes over the Conterminous USA (ConUS) were used to derive the lightning-forecast regression equations:

a. The NOAA North American Regional Reanalysis (NARR) containing 3-D fields from 00:00, 03:00, 06:00, 09:00, 12:00, 15:00, 18:00, and 21:00 UTC for the period January 1990 - August 2018 were downloaded in GRIB2 format and archived on the RMC machines.

b. VAISALA NLDN (U.S.) lightning data for the lower 48 states were obtained through BLM for the period from January 1990 through August 2018.

c. Global Forecast System (GFS) 3-h forecast data fields in GRIB2 format (0.25 x 0.25 degree through 10 days) were archived for the peak fire season (April – September) of 2018 for later testing and verification of the lightning forecast equations.

2.2. Statistical Methodology and Computational Procedures

Our statistical method involved the use of Principal Component Analysis (PCA) with orthogonal rotation to reduce the large cohort of 3-D meteorological driving variables from the NARR dataset to a smaller subset of statistically significant lightning predictors, which were then subjected to logistic regression to produce equations for calculating the probably of *one or more* and *10 or more* cloud-to-ground (CG) flashes. The resulting set of equations can easily be applied to similar forecast fields generated by GFS, WRF or other numerical weather prediction

Table 1. Datasets utilized in the lightning-forecast model development and testing.

Dataset Name & Description	Record Length (years)	Binary Size (Gbit)
North American Regional Reanalysis (NARR): 3-D gridded historical weather dataset provided by NOAA.32-km horizontal resolution interpolated to 20-km resolution for ConUS with a 3-hour time step.	29	3,000
VAISALA NLDN (U.S.): Lightning dataset gridded to 3-hour time steps at 20-km resolution.	29	400
NCEP GFS Forecast Fields: (3-hour time steps from 0 to 240 hours/10 days): 3-D grids of 0.25 x 0.25 degree resolution interpolated to 20-km resolution. Will be used for forecast testing and verification purposes and 2019 forecasts.	1	8,000

models to generate CG forecast probabilities. The "R" statistical package, an open-source software, was utilized to perform PCA and the logistic regressions required to derive the final logistic equations. Resampling the original datasets to a common 20-km resolution national grid was done using the GEMPAK (GEneral Meteorology PAcKage) software jointly developed by NASA and Unidata. The data processing went through the following 8-step procedure:

i. Download and archive 3-D 32-km resolution NARR meteorological fields at 3-hour increments for a period of 29 years (i.e. from 1990 through 2018).

ii. Re-grid NARR fields to 20-km resolution using GEMPAK.

iii. Divide ConUS into climatological Regions. Previous research has shown that deriving region-specific sets of equations significantly improves the overall skill of the lightning forecast model compared to a single national set of equations. Taking into account topography, fuels, climate, and moisture sources, the lower 48 states were aggregated (divided) into 10 Regions depicted in Fig. 1.

iv. Create weather climatology for each ConUS Region using re-gridded 20-km resolution NARR data for 27 years.

v. Create lightning climatology for each ConUS Region to match the NARR climatology using VAISALA NLDN detection data for 27 years.

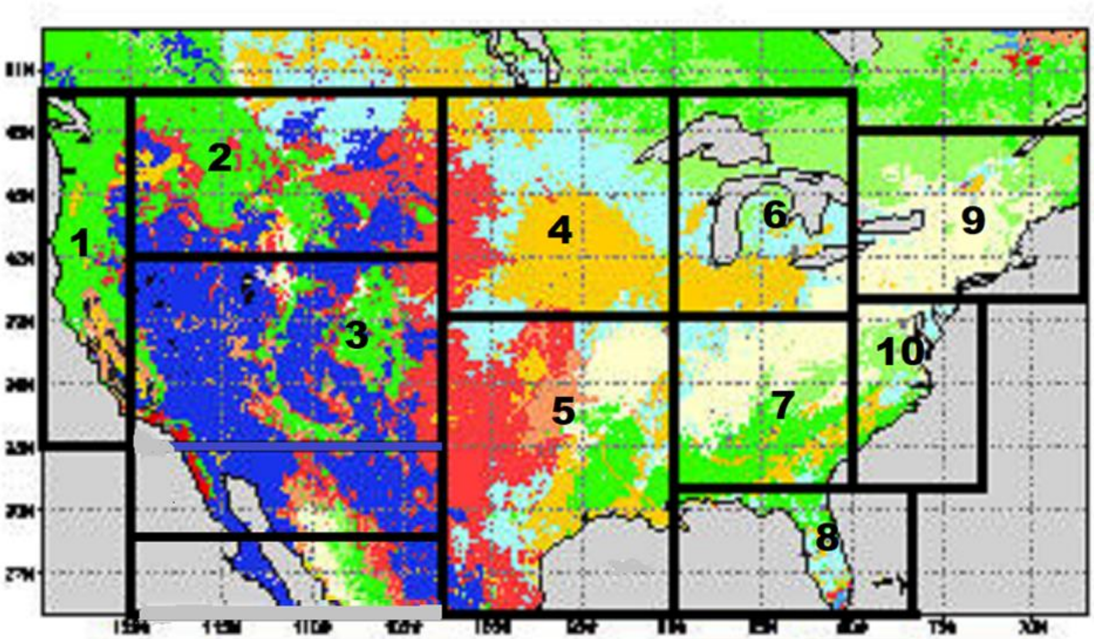


Figure 1. Climatological regions employed in the derivation of lightning forecast equations.

vi. Perform PCA with Varimax/orthogonal rotation on NARR and VAISALA climatological data employing 611 potential meteorological drivers to identify a subset of the best lightning predictors (Fig. 2). This resulted in 13 Principal Components (PCs) used as predictors in all 10 Regions and for each 3-hour period. The predictors referred to NARR vertical pressure levels from 700 mb through 100 mb at 50-mb increment (Fig. 3). The 13 PCs (predictors) explain about 61% of the total variance of CG flashes in all regions and time periods. Temperature & height fields emerged as the strongest predictor among the 13 PCs explaining 14.6% of the observed CG variance.

vii. Create input files and processing scripts for the logistic regression using the "R" statistical software.

viii. Run "R" logistic regression scripts using the top 13 Principal Components (PCs) to derive predictive equations for calculating probabilities of one or more flashes for each time period in every Region. Figure 4 depicts the general mathematical form of logistic equations.

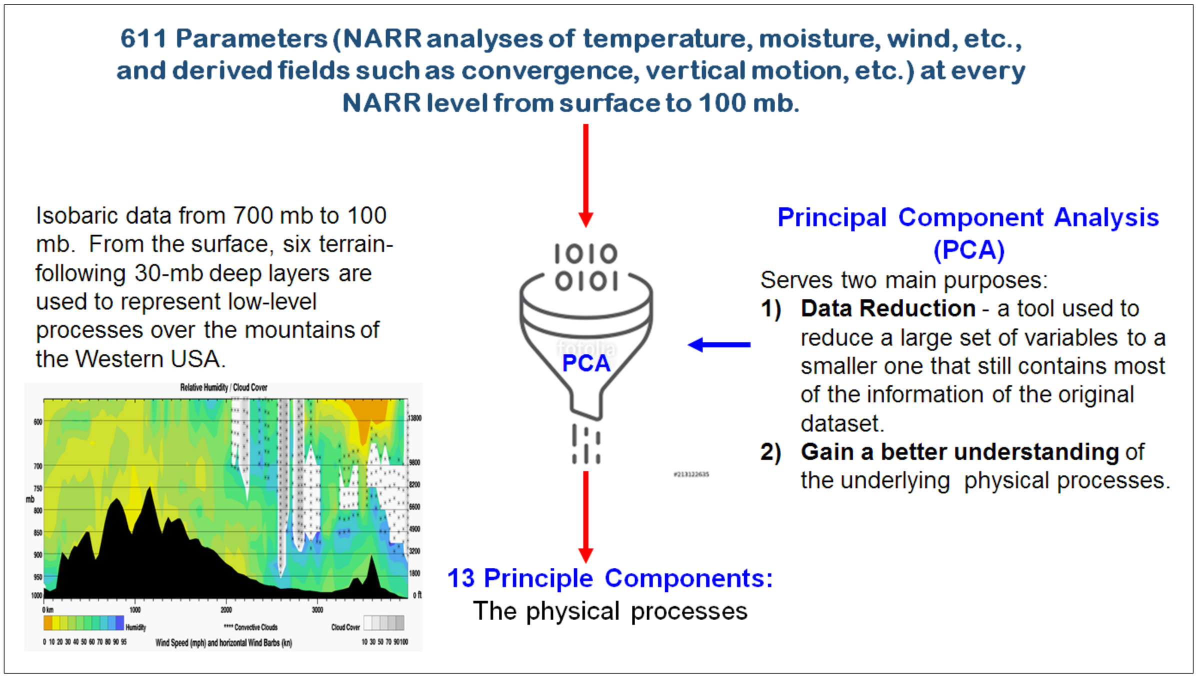


Figure 2. Schematic representation of the Principal Component Analysis (PCA) employed in this study to identify the most significant lightning predictors.

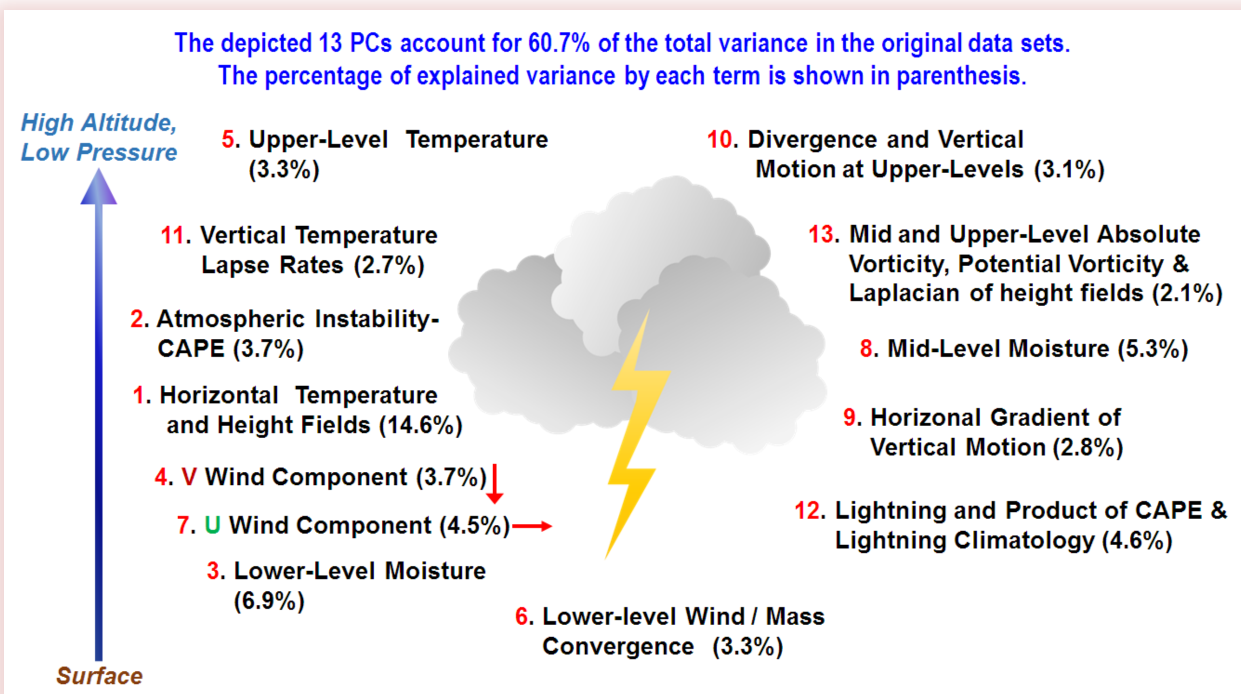


Figure 3. The 13 main lightning predictors (principal components) generated by PCA and the amount of variance of CG flashes explained by them.

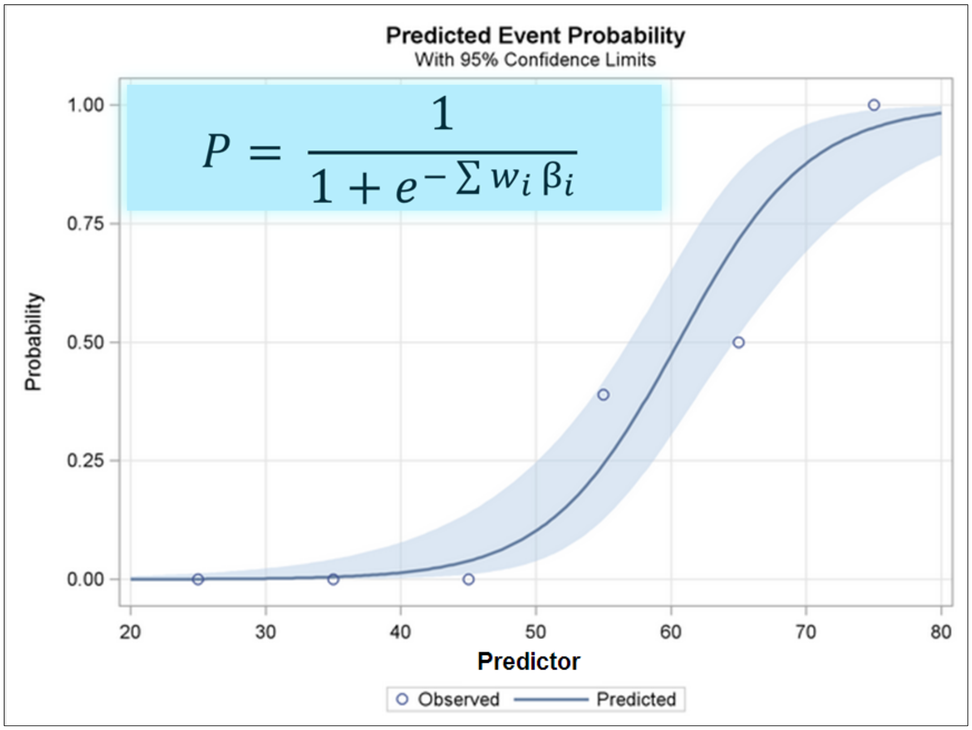


Figure 4. General form of the logistic equations used to calculate the probabilities of one or more CG flashes derived from NARR and VAISALA NLDN data: β_i is the *i*th predictor (principal component) and w_i is the weight (regression coefficient) for that predictor.

3. **Results**

Logistic equations for predicting probabilities of CG flashes were parameterized from a 27-year long record of NARR and VAISALA gridded climatologies for each forecast period of every month in every Region. Each month of the year is described by 8 logistic equations per Region, i.e. one equation for every 3-hour period of the average 24-hour diurnal cycle for that month (the diurnal averages were computed across 27 years). Hence, the number of predictive equations derived across all Regions is 8 x 10 = 80 per month. Due to insufficient CG lightning data in winter months, the quarter December through February was represented by a single diurnal set of 3-h logistic equations. **Thus, the total number of predictive logistic equations for all months of the year in all 10 Regions is about 720.** Table 2 provides an example of regression coefficients (weights) computed for the 13 principal components (predictors) shown in Fig. 3 for June (at 0:00 h UTC) in Region 3 (Southwest).

When run operationally, the lightning-prediction equations coded in special "R" output files can be combined with 3-D meteorological fields produced by either NARR or numerical weather-prediction models to generate lightning probability forecasts across ConUS at 20-km resolution out to 7-10 days.

Table 2. Sample logistic regression coefficients (w_i values in Fig. 4) for June at 0:00 h UTC in Region 3 (Southwestern US).

Predictor	Coefficient
Intercept	-5.32528
Horizontal Temperature and Height Fields	0.45362
Atmospheric Instability including CAPE	1.16505
Low-Level Moisture	1.44811
"V" Wind Components	-0.04265
Upper-Level Temperatures	0.15643
Low-Level Wind/Mass Convergence	-0.26432
"U" Wind Component	-0.37245
Mid-level Moisture	1.70487
Horizontal Gradient of Vertical Motion	0.16518
Divergence and Vertical Motion at Upper Levels	-0.33173
Vertical Temperature Lapse Rate	0.62042
Lightning Climatology & Product of Climatology and CAPE	1.16969
Absolute Vorticity, Potential Vorticity, Laplacian of Heights	-0.17543

3.1. Initial Model Verification

Although a comprehensive verification of the new lightning forecast model is planned for 2019, we performed a limited comparison of predicted CG-flash probabilities to independent lightning observations by VAISALA employing 2018 NARR meteorological fields as drivers that have not been used in the logistic regressions. Thus, this verification has the strength of a model validation. Figures 5 through 7 show predicted probabilities of one or more flashes (color contours) overlapped on the actual observed CG flashes by the VAISALA network (yellow data clusters). Each figure illustrates several 3-h periods from the 20th and 21st days of June, July and August 2018, respectively. A higher modeled probability implies more CG flashes, which is indeed what we observe on the maps. Note that measured CG flashes tend to cluster in areas with modeled probabilities of 30% or higher. Figure 8 shows a Reliability Diagram comparing forecast lightning probabilities to observed frequencies of CG flashes for three 48-hour periods in June, July and August 2018 over the entire ConUS. These preliminary verification results indicate a good skill of the new lightning forecast model that could contribute toward improving the 7-Day Fire Potential Outlook Product of Predictive Services.

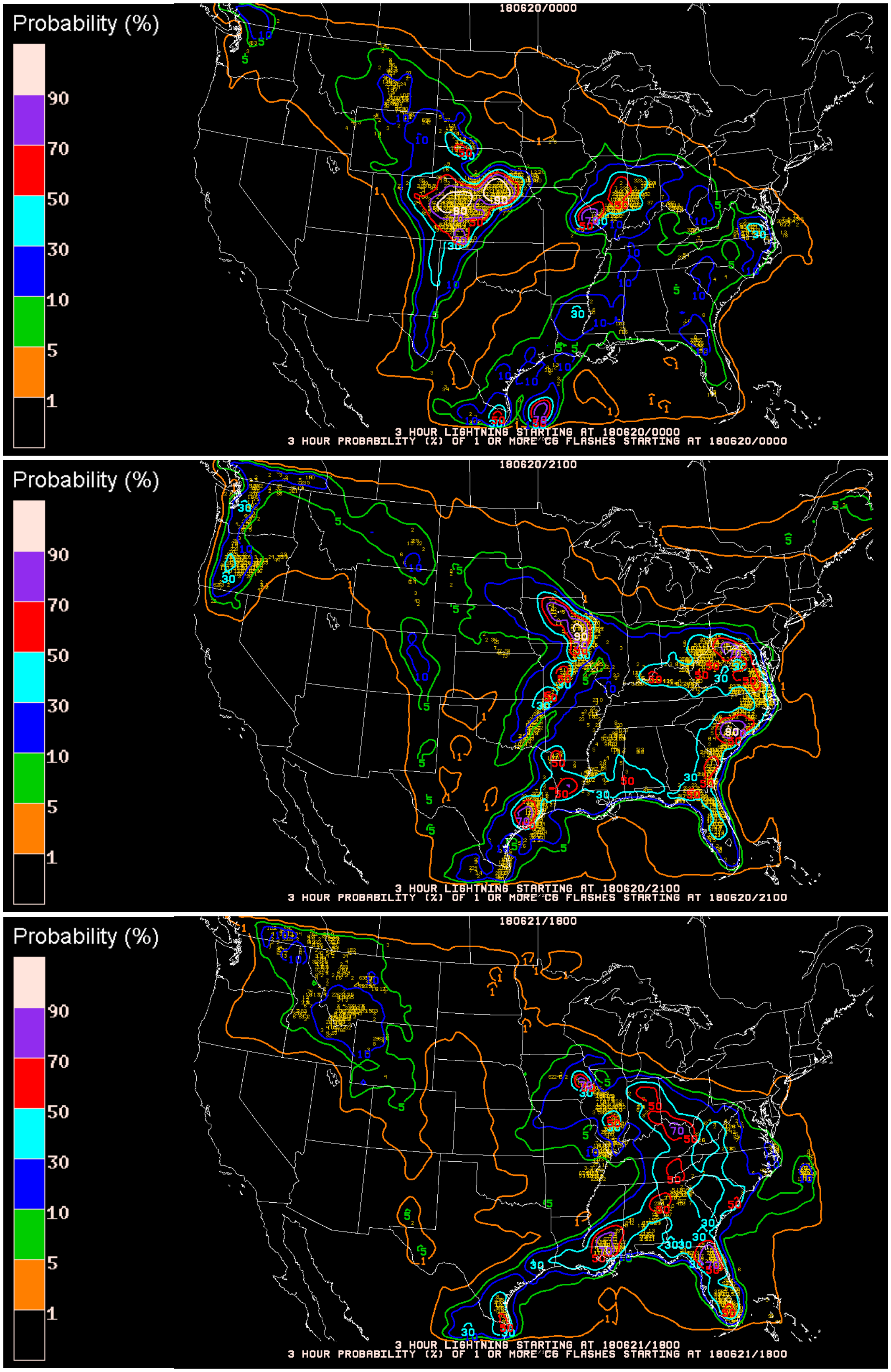


Figure 5. Modeled probabilities of one or more CG flashes (contours) overlaid on observed CG flashes (yellow clusters) for *June 20-21*, 2018 using *independent* NARR fields and VAISALA lightning data.

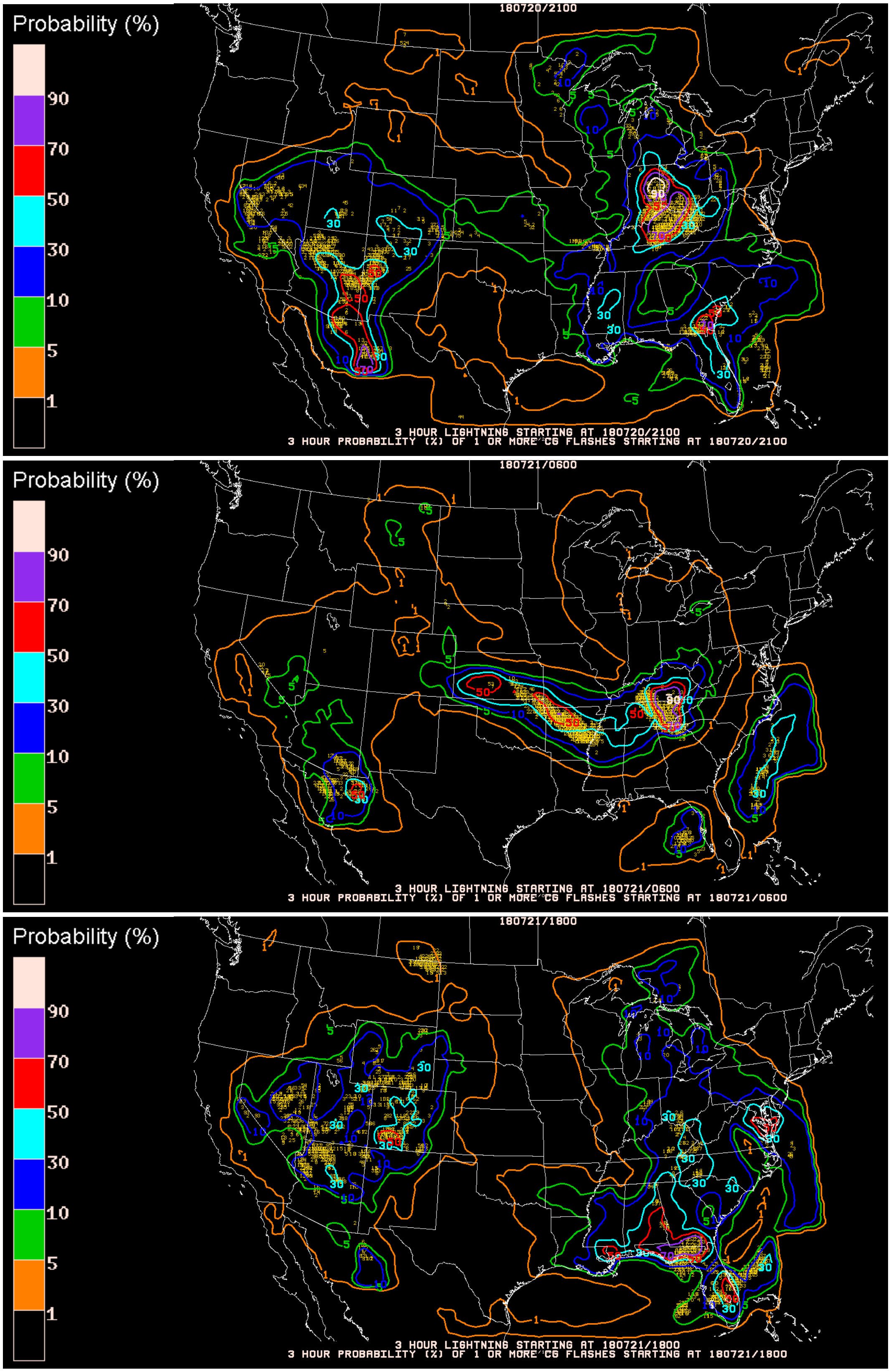


Figure 6. Modeled probabilities of one or more CG flashes (contours) overlaid on observed CG flashes (yellow clusters) for *July 20-21*, 2018 using *independent* NARR fields and VAISALA lightning data.

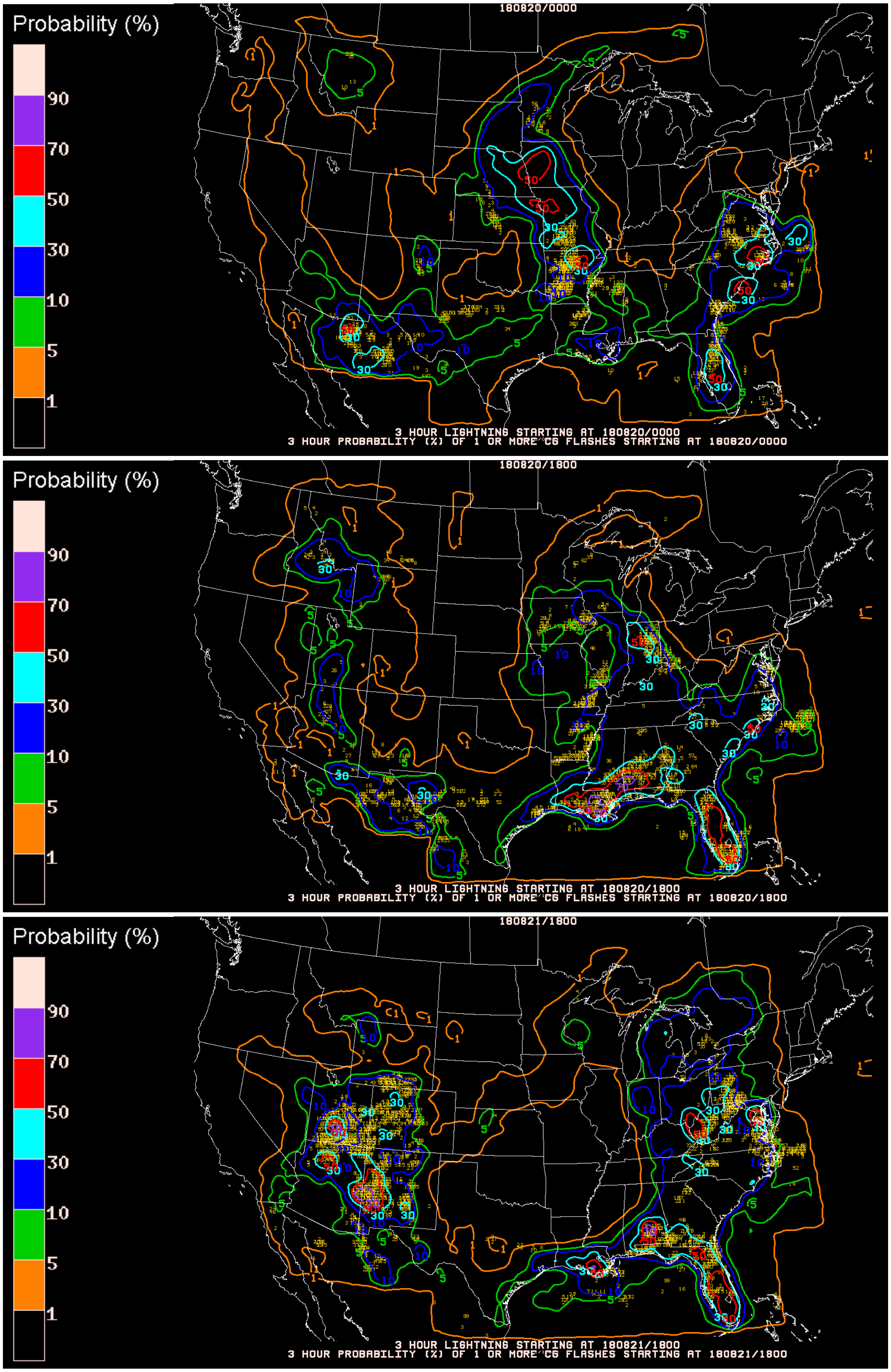


Figure 7. Modeled probabilities of one or more CG flashes (contours) overlaid on observed CG flashes (yellow clusters) for *August 20-21*, 2018 using *independent* NARR fields and VAISALA lightning data.

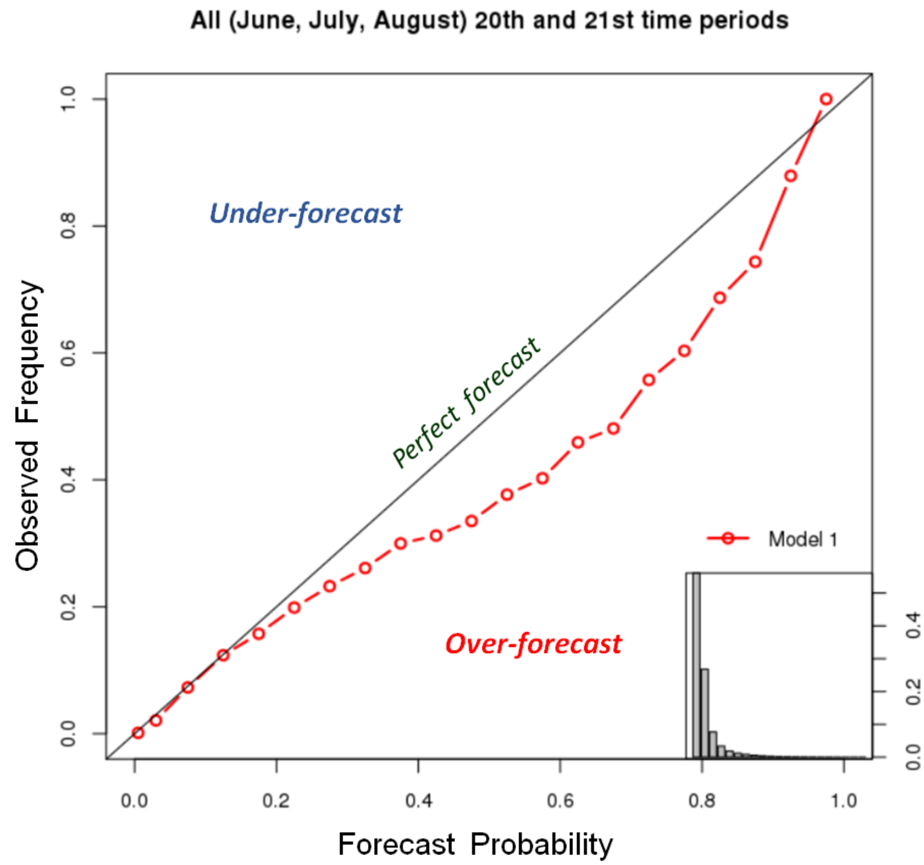


Figure 8. Reliability diagram comparing forecast lightning probabilities to observed frequencies of CG flashes for ConUS over three 48-hour periods between June 20 and August 21, 2018 using independent NARR and VAISALA data.

4. Conclusion

This is the only model capable of predicting lightning probabilities on a 20-km resolution grid over ConUS derived from the longest climatological record of 27 years available. Initial verification results indicate a significant model skill in forecasting the chance of one or more CG flashes across ConUS. The 3-h temporal resolution of the model allows a seamless integration with standard outputs from NWS numerical weather prediction models such as GFS and NAM.

5. Future Work

Our future work will focus on producing predictive numerical algorithms that *translate* forecast changes of weather for the next 7-10 days into key fire activity indicators such as likelihood of ignition, potential fire growth etc. To insure compatibility between various indicators, we propose quantifying all of them as *probabilities*. This will allow combining different indicators into new indices quantifying the fire danger and resource-allocation needs on a uniform national grid.

5.1. Proposed Work for 2019:

We envision to conduct the following R&D tasks during Phase 2 of the Joint Venture Agreement in 2019:

a. Complete the verification of the lightning forecast algorithm for the lower 48 states.

b. Compile a historical lightning data set for Alaska.

c. Derive logistic equations for predicting fire-ignition probabilities out to 7-10 days as a function of weather & fuel conditions on a 20-km resolution grid for the lower 48 states.

Tasks *a* and *b* represent a continuation of the research performed in 2017/2018. **However, Task *c* is new and requires coordination with and input from PS meteorologists and the PS Oversight Group.** With the intent to initiate a dialog on this topic between RMC and PS Oversight Group we propose the following plan for moving forward on this task.

Proposal for tackling Task *c*. The historical NARR data will be used to compute a variety of weather indices based on fields of temperature, relative humidity, dew point, wind speed, 24-h precipitation, atmospheric stability and others. We plan to utilize a land-cover dataset defining relevant fuel types such as the NASA/NCEP North American Land Data Assimilation System or Matt Jolly's high-resolution Fuel Map. The weather-based fire indices will include but not be limited to NFDRS parameters (such as ERC) and NOAA's Evaporative Demand Drought Index (EDDI). **We seek input from meteorologists at each GACC regarding *potential predictors* of fire-ignition probabilities that they have identified and successfully used in the past.** Based on this input and our own analysis, we will compile a comprehensive list of relevant predictors. These will then be run through PCA along with lighting climatologies produced in Phase 1 of the Joint Venture Agreement and historical fire occurrences from the K.C. Short data set. PCA will yield the best set of predictors for fire-ignition probabilities at each 20-km grid point across the lower 48 stated. Finally, the predictors at every grid point identified by PCA will be harnessed to derive equations for fire-ignition probabilities using a logistic regression. The general form of the logistic probability equation is:

$$P = \frac{1}{1 + e^{-x}}$$

where $0 \leq P \leq 1$ is the probability for fire ignition over a period of 7-10 days and $x = \sum w_i \, p_i$ is a composite term consisting of a linear combination (sum) of various predictors (p_i) multiplied by their weights (w_i) calculated as regression coefficients specific to each grid point. We will perform the analysis using *all* (natural and human-caused) fires present in the K.C. Short data set.

5.2. Proposed Work for 2020:

Develop Equations for forecasting the probabilities of fire growth beyond certain threshold sizes applicable to a 20-km ConUS grid and derived via multiple logistic regression using data from the following sources:

- Karen Short's Spatial Wildfire Dataset: 1992 - 2017;
- MODIS Active Fire and Burned Area Products Dataset;
- NCEP North American Regional Reanalysis (NARR) Dataset: 1992-2018;
- Gridded Fuel Types from the NASA/NCEP North American Land Data Assimilation System;
- Gridded weather-based drought and fire-danger indices such as EDDI, ERC etc.

Again, we are actively seeking *input* from all GACCs and the PS Oversight Group on both the type of fire-activity indicators to be quantified and their predictors to be considered in our analyses.