

Probabilistic Fire Danger Forecasting: A Framework for Week-Two Forecasts Using Statistical Postprocessing Techniques and the Global ECMWF Fire Forecast System (GEFF)

ROCHELLE P. WORSNOP,^{a,b} MICHAEL SCHEUERER,^c FRANCESCA DI GIUSEPPE,^d CHRISTOPHER BARNARD,^d
THOMAS M. HAMILL,^b AND CLAUDIA VITOLO^d

^a Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, Colorado

^b NOAA/ESRL, Physical Sciences Laboratory, Boulder, Colorado

^c Norwegian Computing Center, Oslo, Norway

^d European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 17 May 2021, in final form 19 August 2021)

ABSTRACT: Wildfire guidance two weeks ahead is needed for strategic planning of fire mitigation and suppression. However, fire forecasts driven by meteorological forecasts from numerical weather prediction models inherently suffer from systematic biases. This study uses several statistical-postprocessing methods to correct these biases and increase the skill of ensemble fire forecasts over the contiguous United States 8–14 days ahead. We train and validate the postprocessing models on 20 years of European Centre for Medium-Range Weather Forecasts (ECMWF) reforecasts and ERA5 reanalysis data for 11 meteorological variables related to fire, such as surface temperature, wind speed, relative humidity, cloud cover, and precipitation. The calibrated variables are then input to the Global ECMWF Fire Forecast (GEFF) system to produce probabilistic forecasts of daily fire indicators, which characterize the relationships between fuels, weather, and topography. Skill scores show that the postprocessed forecasts overall have greater positive skill at days 8–14 relative to raw and climatological forecasts. It is shown that the postprocessed forecasts are more reliable at predicting above- and below-normal probabilities of various fire indicators than the raw forecasts and that the greatest skill for days 8–14 is achieved by aggregating forecast days together.

KEYWORDS: Forest fires; Wildfires; Bias; Statistical techniques; Uncertainty; Ensembles; Forecast verification/skill; Hindcasts; Numerical weather prediction/forecasting; Probability forecasts/models/distribution; Reanalysis data


1. Introduction

Global wildfires continue to break records year after year in scale, duration, and impact on human life, ecosystems, and property (Jolly et al. 2015; Wang et al. 2021). In 2020 alone, wildfires in the western United States burned 75% more area than expected in an average year and are expected to total hundreds of billions of dollars in direct and indirect costs (Fu et al. 2021). Enhanced wildfire size, duration, and impact are in part due to human behavior: 1) increased residential developments in the wildland–urban interface (Radeloff et al. 2018) lead to more possibilities of human-caused wildfires and more loss of life and property and 2) past logging and legacy forest-fire management practices have led to a buildup of fuels (Parks et al. 2015). However, enhancement also comes from changes in the atmosphere and responding vegetation as a result of anthropogenic climate change. Changes in regional climate such as delayed onset of wet seasons, earlier snowmelt, increased size of high pressure blocking patterns, increased aridity, and hotter temperatures (Flannigan et al. 2013, 2000; Westerling et al. 2006; Jolly et al. 2015; Xu et al. 2020; Schoennagel et al. 2017;

Williams et al. 2019; Nabizadeh et al. 2019 and references therein) all contribute to prime fire conditions. Additionally, wildfire risk is exacerbated by “hot droughts” or the concurrence of droughts and heat waves, which are predicted to worsen across water-limited regions in the United States as a result of the warming climate (Cheng et al. 2019).

While wildfires can be incredibly destructive, they play an integral role in the Earth system’s patterns and processes (Bowman et al. 2009) and their impacts to societies are ultimately inevitable. One of many ways to create community resilience to wildfires is with skillful forecasts of fire indicators at various forecast horizons. The U.S. National Weather Service Storm Prediction Center provides fire-weather outlooks for day 1, day 2, and combined days 3–8 in the form of written synoptic discussions and mapped regions of categorical fire risk given weather and antecedent fuel conditions (SPC 2021). The forecasters use deterministic fire indicators such as those within the National Fire Danger Rating System (NFDRS; described in section 2) as one of their tools to assess current fire conditions and make predictions.

Unfortunately, most fire indicators used in the United States are calculated from observations and guidance have mostly been produced either for real-time or one-day ahead. Fortunately, several studies (Mölders 2010; Di Giuseppe et al. 2016, 2020; Worsnop et al. 2020) have shown that weather forecasts output from numerical weather prediction (NWP) models can fill in the spatial gaps where observations do not exist and can extend fire-indicator guidance to longer lead times. Using this approach,

 Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-21-0075.s1>.

Corresponding author: Rochelle P. Worsnop, Rochelle.Worsnop@noaa.gov

useful forecasts can be extended up to two weeks ahead, and, in some cases, out to seasonal time scales (Bedia et al. 2018). Even further, ensemble prediction systems (EPSs) in which a finite number of varying forecast scenarios are generated for a given forecast period can be used to quantify the uncertainty associated with a forecast. Different ensemble members are commonly generated by making small perturbations to the initial conditions or physics schemes used in the NWP model. While the ensemble enables valuable probabilistic prediction, EPSs often suffer from systematic biases and dispersion errors (Stauffer et al. 2017).

To generate the most skillful probabilistic fire forecasts at the extended range (i.e., days 8–14), statistical postprocessing is necessary (Worsnop et al. 2020). The goals of statistically postprocessing ensemble forecasts are threefold: 1) remove systematic biases inherent in NWP forecasts, 2) provide probabilistic forecasts with reliable ensemble spread, and 3) maximize sharpness so that forecast prediction intervals are shorter than that of climatology (i.e., on average, the forecast should be at least as skillful as this benchmark). This concept is referred to in the literature as *maximizing the sharpness of the predictive distributions subject to calibration* (Gneiting and Raftery 2007).

The statistical postprocessing methods presented herein adjust the raw ensemble forecast, in an effort to improve skill and quantify uncertainty, to account for discrepancies between past forecast–observation pairs. All of the univariate postprocessing methods used in this study are parametric and therefore relate characteristics of the raw ensemble to parameters of the resulting calibrated predictive distribution over a training period (i.e., distributional regression models).

Once the meteorological forecasts are postprocessed (i.e., calibrated), they can be used to generate postprocessed fire indicators. Calibrated meteorological variables have previously improved extended-range forecasts of the hot–dry–windy index (HDWI; Srock et al. 2018; Worsnop et al. 2020), an index based solely on meteorological conditions. While it is valuable for a forecaster to assess weather-driven fire, a fire will not burn or spread unless the vegetation is receptive to fire. For this reason and because the fire community has requested operational forecasts (at the Climate Prediction Center) that incorporate fuels information, we input the calibrated ensemble weather forecasts to the Global ECMWF Fire Forecast (GEFF) System (Di Giuseppe et al. 2016). The outputs from the GEFF provide an assessment of broad-scale fire danger based on interactions among atmospheric conditions, topography, and the state of vegetation/fuels.

The remainder of the paper provides an overview of the GEFF system in section 2 with a focus on fire indicators commonly used in the contiguous U.S. (CONUS) domain. The NWP reforecasts and reanalysis data used for training and verification are discussed in section 3 followed by the data preprocessing steps in section 4. The postprocessing steps are detailed in section 5. Verification of the skill of the postprocessed fire-indicator forecasts compared to standard benchmarks is presented in section 6. Conclusions and a discussion of how this framework will be used in operations are covered in section 7.

2. The Global ECMWF Fire Forecast (GEFF) system

Figure 1 shows a conceptual diagram of the steps we take to calculate the components of fire danger rating systems given meteorological inputs of reanalysis or reforecast data (discussed in section 3). One tool we use is the GEFF. The open-source GEFF system (<https://github.com/ecmwf-projects/geff>) was developed by ECMWF and the European Joint Research Centre (JRC) to provide daily operational forecasts of three fire danger rating systems for the European Forest Fire Information System (Camia et al. 2006). ECMWF generates forecasts up to 10 days ahead by providing the GEFF with atmospheric forcings at 9 (19)-km grid spacing from the HRES (ensemble) run of the ECMWF Integrated Forecast System (IFS). While these forecasts are created to monitor forest fire danger in European Union countries, GEFF forecasts can be produced for anywhere in the world. Additionally, the GEFF can use meteorological inputs from any NWP model, reanalysis, or observations datasets.

The flexibility of the GEFF enables us to investigate how our postprocessing techniques impact the skill of fire-indicator forecasts over the CONUS domain by comparing the outputs from the GEFF system generated with raw and postprocessed forcings from a NWP model to those generated with reanalysis data. Data used for this study are discussed in section 3.

The GEFF framework is described in detail in (Di Giuseppe et al. 2016) and works similarly here. In general, for a given location, the GEFF system uses three tiers of data representing different temporal scales (see Fig. 1). Tier 1 uses climatological inputs that remain constant throughout the forecast cycle. Slope and fire-fuels information such as fuel model, vegetation cover and stage are in this tier as well as mean annual precipitation. We used the same tier 1 data discussed in detail in the appendix of (Di Giuseppe et al. 2016). Data in tier 2 and tier 3 must be calculated with our meteorological forcings at daily (i.e., minimum, maximum, or 24-h accumulations) and local-noon time scales, respectively. Table 1 lists the inputs used in tiers 2 and 3.

These climate, daily, and local-noon forcings are then passed to the GEFF system, which performs various empirical and physics-based equations that relate climate and meteorological conditions to fuel response and subsequently forest-fire behaviors. Conveniently, the GEFF system can use these inputs to calculate components of three, well-established fire-rating systems: the Australian McArthur's Forest Fire Danger Meter (MARK 5; McArthur 1966), the Canadian Fire Weather Index (FWI; Van Wagner 1987, 1974), and the U.S. Forest Service National Fire Danger Rating System (NFDRS; Cohen and Deeming 1985; Bradshaw et al. 1983; Schlobohm et al. 2002).

All three of these rating systems use a combination of factors relating to weather, fuels, and topography, which affect fire ignition, spread, and consequently the difficulty to control a wildfire if one started. Daily calculations to estimate today and tomorrow's fire danger are typically made with in situ meteorological and fuels data from discrete observation stations. The estimates are then interpolated to areas between observation sites using techniques such as inverse distance weighting. Because of the availability of local measurements of the fuel and

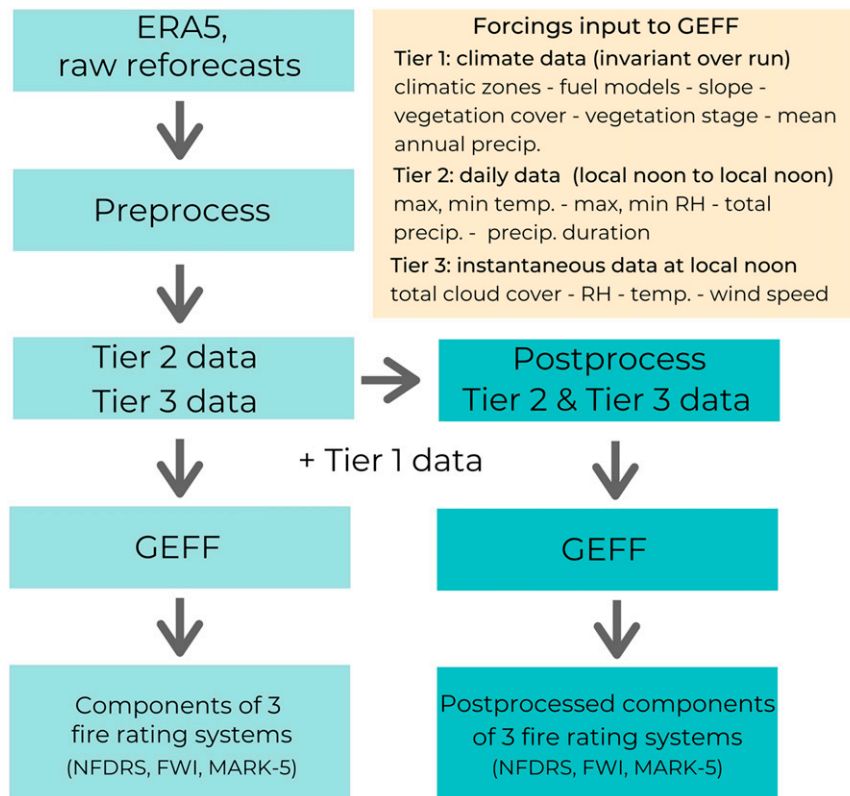


FIG. 1. Conceptual diagram of the steps we take to generate “observed” (using ERA5 data) and forecasted components of the fire-rating systems output from the GEFF system. The darker teal boxes indicate the steps needed to generate calibrated ensemble forecasts. The orange box describes the three different tiers of data input to the GEFF system. The results in this paper focus on the components of the National Fire Danger Rating System (NFDRS).

meteorological conditions, estimates of today or tomorrow’s fire danger at station locations or at locations between a close network of observation stations are likely more precise than estimates using NWP forecasts as inputs. However, the gridded forecasts used herein provide gridpoint-by-gridpoint estimates across CONUS and enable fire forecasting up to two weeks ahead.

National fire danger rating system

The focus domain of this study is the CONUS, so we will concentrate on the NFDRS, which is most commonly used to assess fire danger across the CONUS. The NFDRS is used by federal and state agencies to make strategic decisions regarding forest fires. Visitors of U.S. Forest Service lands may recognize the NFDRS qualitative class ratings from low to very extreme displayed on “Smokey Bear” signs. These ratings describe the relative severity of the fire danger situation and are derived from the quantitative components and indicators of the NFDRS. The key components and indicators of the NFDRS used to derive these ratings are:

1) Spread component (SC): an open-ended dimensionless quantity of the forward spread of the head of the fire that is dependent on factors such as wind speed, slope, and moisture of fine fuels and live woody fuels.

2) Energy release component (ERC): an open-ended quantity reflecting the available energy (i.e., potential heat release) per unit area (BTU ft^{-2}) within the flaming zone of the head of the fire through the burning of live and dead fuels.

3) Burning index (BI): an open-ended indicator of the potential effort required for fire suppression based on how much heat would be released from the fire and how fast it would spread

TABLE 1. Preprocessed meteorological variables input to the GEFF system.

Tier	Variables	Abbreviation	Units
2	Min 24-h temperature	$t_{\{24h, \min\}}$	K
	Max 24-h temperature	$t_{\{24h, \max\}}$	K
	Min 24-h relative humidity	$RH_{\{24h, \min\}}$	%
	Max 24-h relative humidity	$RH_{\{24h, \max\}}$	%
	24-h precipitation accumulation	tp_{24h}	mm
	24-h precipitation duration	dp_{24h}	h
3	Local-noon total cloud cover	tcc_{noon}	Area fraction
	Local-noon relative humidity	RH_{noon}	%
	Local-noon temperature	t_{noon}	K
	Local-noon wind speed	u_{noon}	m s^{-1}
	Local-noon snow cover mask	$sc_{\{m, \text{noon}\}}$	Binary

in a worst-case scenario. It is expressed in feet multiplied by 10 and is a combination of the SC and the ERC.

- 4) Ignition component (IC): the probability from 0 to 100 that a firebrand in the presence of receptive fuels will ignite a fire that would require suppression activities.

Operational national maps of the adjective class ratings are predominantly derived from the BI followed by the ERC. The classes are defined by percentiles of the BI and ERC climatologies where the breakpoints are determined from historical relationships between the BI or ERC and observed fires within a given area. These breakpoints vary (e.g., 80th versus 97th percentile as the starting breakpoint) depending on agencies and the types of decisions that need to be made for a given area and situation. Due to the variable nature of these breakpoints, we will focus our analysis on the quantitative components and indices of the NFDRS. However, in operations, agencies and forecasters could convert these outputs to classes for decision making. The FWI and MARK5 are similar to the NFDRS and mostly differ in their use of fuel models and types. Di Giuseppe et al. (2016) provides an overview of their differences.

3. Reforecasts and reanalysis data

The meteorological forecasts input to the GEF system, either raw or postprocessed, are based on 11-member ensemble reforecasts (i.e., decades of past forecasts reran with the same model version) from the Cycle45r1 version (ECMWF 2018) of the ECMWF Integrated Forecast System (IFS). A total of 20 years of reforecasts were produced from the 0000 UTC initialization run for every Monday and Thursday between 7 June 2018 and 6 June 2019. Therefore, our reforecast dataset comprises 104 initialization forecasts and 2080 (20×104) total reforecasts at each lead time. We downloaded 6-hourly reforecasts with lead times out to 15 days over CONUS with a regular latitude–longitude grid with 0.25° grid spacing from the ECMWF MARS archive system. The domain covers 23° – 53° N latitude and 126° – 65° W longitude. We retrieved reforecasts for the meteorological variables shown in Table 2.

We use reanalysis data to calibrate and verify the postprocessed forecasts of meteorological variables and the subsequent fire indicators output from the GEF system. From the Copernicus Climate Change Service Climate Data Store (C3S 2018), we retrieved hourly, ERA-5 reanalysis data (Hersbach 2000; Hersbach et al. 2018) for the variables in Table 2 with 0.25° grid spacing on the same domain as the reforecasts for all dates between 6 June 1998 and 6 June 2018 (i.e., the first and last date in the reforecast dataset) plus 15 extra days to match the forecast horizon out to 15 days. We also retrieved data for the date before the first reforecast date (5 June 1998), which was needed to perform the preprocessing steps discussed in section 4.

4. Preprocessing steps

The tier 2 inputs to the GEF system are based on daily minimum, maximum, and accumulated quantities over the preceding 24-h period since the previous day's local noon.

TABLE 2. Reforecast and reanalysis variables downloaded and used to calculate inputs to the GEF system.

Variables	Abbreviation	Units
2-m temperature	t2m	K
2-m dewpoint temperature	t _d 2m	K
10-m <i>u</i> -wind speed	u10m	m s^{-1}
component		
10-m <i>v</i> -wind speed	v10m	m s^{-1}
component		
Total cloud cover	tcc	Area fraction
Total precipitation	tp	m
Snow depth	sd	m of water equivalent

The tier 3 data are based on instantaneous quantities at local noon today. This local noon adjustment is part of the requirements to calculate the components of the NFDRS and is considered the time with the most favorable wildfire conditions.

We determine the UTC corresponding to local noon at each grid point using the TimeZoneFinder Python package (<https://pypi.org/project/timezonefinder/>). Over the CONUS domain, local noon ranges from 1600 to 1900 UTC (1700–2000 UTC) within four possible daylight (standard) time zones. Since the ERA5 reanalysis and ECMWF reforecasts start at 0000 UTC and local noon ranges anywhere from 1600 to 2000 UTC, 24-h accumulations and minimum and maximum quantities must use some of the previous day's reanalysis or forecast data along with today's data to form a 24-h period. Because the ERA5 reanalysis is hourly, the tier 2 data are simply the minimum, maximum, and accumulated quantities over this 24-h period preceding local noon and the tier 3 data are the instantaneous values at local noon. Since the reforecasts are output every 6 h, local noon does not always align with the 6-hourly forecast output (i.e., 0000, 0600, 1200, 1800 UTC). For these situations, we use weighted linear interpolation to estimate the meteorological variables at the local noon time to generate the values at the start and end of the 24-h period. The resulting (preprocessed) tier 2 and tier 3 variables input to the GEF are shown in Table 1.

5. Postprocessing methods

This section describes the different univariate and multivariate statistical methods used to postprocess the meteorological variables input to the GEF system. The GEF system does not rely on any particular postprocessing algorithm; it can accept any meteorological input. Here, we used the censored, shifted Gamma distribution model (CSGD; Scheuerer and Hamill 2015) to postprocess accumulated precipitation and an ensemble model output statistics (EMOS; Gneiting et al. 2005) approach to postprocess temperature and wind speed variables. For the remaining variables—total cloud cover, snow cover mask, precipitation duration, and relative humidity variables—we use variants of the logistic regression method (Hemri et al. 2016; Wilks 2011).

a. Training and verification periods

We use cross-validation to estimate and verify the parameters of the various regression equations in the methods below.

The training set consists of all reforecast–observation pairs from a defined time window for a given grid point and lead time for all reforecast years except for the verification year (i.e., 19 total training years). Verification is performed with reforecast–observation pairs from the left-out verification year. The lengths of the windows are described below and vary from 1 to 3 months depending on the variable and method.

b. Univariate postprocessing methods

1) PRECIPITATION ACCUMULATION

Precipitation accumulation is a unique variable to postprocess, namely, because of the three features outlined in [Scheuerer and Hamill \(2015\)](#) and reiterated here: 1) the distribution has a combination of discrete and continuous characteristics with a point mass for exactly zero precipitation, 2) the forecast uncertainty typically increases with precipitation amount, and 3) a large and varied training dataset is required for accurate prediction of rare events such as excessive rainfall amounts.

The CSGD distributional regression model addresses each of these peculiarities and yields a predictive distribution (“censored, shifted gamma distribution”) with three defining parameters related to the mean, standard deviation, and shift of the distribution, respectively. We use the same set of ensemble statistics used in [Scheuerer and Hamill \(2018\)](#): the ensemble mean, the ensemble probability of precipitation, and the ensemble mean absolute difference to relate the raw forecasts to the CSGD parameters. The ensemble statistics are calculated using all forecasts within a weighted spatial neighborhood following the procedures in [Scheuerer and Hamill \(2015\)](#) to account for any displacement errors in the raw forecast. We fit the CSGD parameters using a 91-day window (see [section 5a](#)) centered around the 15th day of each month. We offer more details about the CSGD-model procedures in the supplemental material while a thorough description of the CSGD model, its implementation, and comparisons to alternative postprocessing methods are discussed in [Scheuerer and Hamill \(2015\)](#).

At dry locations, there may not be enough nonzero forecast–observation pairs in the training dataset to avoid overfitting the CSGD model. For these locations, we use ad hoc values determined through exploratory analysis for the parameters of the regression equations in step 3. See supplemental material for details.

The results of this section’s steps are calibrated predictive distributions of precipitation accumulations conditional on statistics of the raw ensemble fitted separately for each month, lead time, and grid point.

2) TEMPERATURE AND WIND SPEED

We briefly describe the univariate postprocessing method used to generate calibrated and sharp marginal predictive distributions of local noon wind speed, and minimum, maximum, and local noon temperature separately for each month, lead time, and location. The cross-validation window size is 28–31 days depending on the month (see [section 5a](#)). The ensemble model output statistics (EMOS; [Gneiting et al. 2005](#)) method uses raw ensemble statistics and links them

to the parameters of a predictive distribution that is well suited to represent the forecast uncertainty about the variable of interest. We use the same variant of the EMOS regression equations that [Worsnop et al. \(2020\)](#) used to postprocess surface temperature and wind speed. These linear equations relate the raw ensemble mean and variance to the mean and variance parameters of the predictive distribution.

Unlike in [Worsnop et al. \(2020\)](#), we do not apply flexible power transformations to the data. While their approach is more flexible and would likely produce the most calibrated and sharp forecasts, the computation is expensive. Given that there are eleven different variables to postprocess before running through the GEFF system, we opt for a more streamlined approach. During exploratory analysis, we tested several candidate distributions (i.e., Gaussian, truncated Gaussian, square root truncated Gaussian, gamma, and truncated logistic) to represent temperature and wind speed observations. We found through continuous ranked probability skill scores (CRPS) and probability integral transform (PIT) histograms that the Gaussian and truncated logistic distributions are well suited to represent 2-m temperature and surface wind speed, respectively over the majority of CONUS, so we processed with this framework.

3) TOTAL CLOUD COVER, PRECIPITATION DURATION, AND RELATIVE HUMIDITY

Total cloud cover, precipitation duration, and minimum, local noon, and maximum relative humidity are doubly bounded variables that can be discretized into a range of possible categories. For these variables, we use the proportional odds logistic regression (POLR; [Hemri et al. 2016](#)) method to predict the conditional probabilities of these categories given a vector of predictors based on raw ensemble statistics. The training and verification periods are the same that were used for temperature and wind speed.

The POLR model, a variant of ordered logistic regression, is a statistical model that can be used to predict conditional probabilities of ordered predictands falling within one of multiple possible categories. The logarithm of odds (i.e., logit) of the probability of the (discretized) predictand y falling within or below a category j , conditional on a vector of predictor variables $\mathbf{x} = (x_1, x_2, x_3, \dots)^T$ is written as

$$\text{logit}[P(y \leq j | \mathbf{x})] = \theta_j - \beta \mathbf{x}, \quad j = 1, \dots, J - 1, \quad (1)$$

where J is the number of categories. Because the probability of falling within or below the last category is always 1, we only build a model for the first $J - 1$ categories. For the POLR model, the slope coefficient vector $\beta = (\beta_1, \beta_2, \beta_3, \dots)$ is the same size as the number of predictors and is the same value for all categories, while the intercept θ_j is estimated separately for each category by selecting the values that maximize the log-likelihood function over the training period. Taking the inverse logit of (1) yields the cumulative probability written as

$$P(y \leq j | \mathbf{x}) = \frac{\exp(\theta_j - \beta \mathbf{x})}{1 + \exp(\theta_j - \beta \mathbf{x})}. \quad (2)$$

We define the predictors $\mathbf{x} = (r_{\text{ENS}}, f_0, f_1)^T$ following one of the POLR model variants used by Hemri et al. (2016). In our study, r_{ENS} is the mean of the M -member ensemble. In Hemri et al. (2016), f_0 and f_1 are the fractions of the ensemble members within the lowest and highest possible category, respectively. We altered f_0 and f_1 to represent the fraction of ensemble members within the lowest and highest category that was actually populated by some ensemble members at the given location, month, and lead time. This adjustment helped account for climatological differences of the various meteorological variables dependent on location and time of year. For example, in a dry location and time of year, tcc_{noon} may never fall into the highest possible category J , so the statistic f_1 is uninformative. However, if we define the highest category depending on the local climatology, there may be some value in knowing that a large number of members all falls into that category. The remainder of this paper will use the modified definition of f_0 and f_1 .

Hemri et al. (2016) tested additional predictor variables such as ensemble variance and an interaction term between the ensemble variance and the ensemble mean's deviation from 0.5. We found that these additional terms did not add substantial skill to the postprocessed forecasts, particularly for longer lead times, which is consistent with the findings by Hemri et al. (2016). Since our main focus of this study is to generate skillful forecasts in the extended range, we perform POLR with the simple set of predictors $\mathbf{x} = (r_{\text{ENS}}, f_0, f_1)^T$.

(i) Defining categories for each meteorological variable

To fit the POLR model, continuous observed meteorological variables (reanalysis data in our case) need to be discretized into J categories binned according to a set of boundaries Ω . Determining the number of categories to bin the data into is a trade-off between bias and variance. Since the POLR model estimates θ_j for each category, more categories lead to a more complex model and the risk of overfitting due to sampling variability. Conversely, too few categories can result in biases when the predicted category probabilities of the predictand are interpolated to a full predictive distribution for the continuous predictand (discussed below) because the approximation of this distribution is too coarse.

In standard observation reports, total cloud cover values are typically reported in nine defined ordered categories called octas ranging from 0 to 8, which represent approximately 1/8th intervals $Z = \{0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 1\}$. We use the same binning thresholds as Hemri et al. (2016) (their Table A1) to map our continuous tcc_{noon} variables to these octas. Specifically, we use $\Omega = \{0.01, 0.1875, 0.3125, 0.4375, 0.5625, 0.6875, 0.8125, 0.99, 1.0\}$ as threshold values that define the upper boundaries of each bin. For example, category 1 (no clouds) falls within $[0, 0.01)$ and category 9 (full cloud cover) falls within $[0.99, 1]$.

We were able to achieve good calibration and skill (according to verification rank histograms and CRPSS, not shown) for precipitation duration ($\text{dp}_{24\text{h}}$) using the same categories after normalizing the duration hours to take on values in $[0, 1]$. However, through exploratory analysis, we increased the number of categories to 22 for local noon, minimum, and maximum relative humidity variables, which resulted in more calibrated

(flat) histograms than when we used the tcc_{noon} categories. This may suggest that relative humidity values are more concentrated into smaller intervals—and require finer discretization—than tcc_{noon} or $\text{dp}_{24\text{h}}$. The 22 categories are defined by upper limits $\Omega = \{0.01; k + 0.05 \text{ for } k = 0, 0.05, 0.10, \dots, 0.95; 0.99; 1.0\}$ after normalizing the relative humidity variables to take on values in $[0, 1]$.

(ii) Generating an ensemble of postprocessed forecasts based on the POLR probability forecasts

The cumulative category probabilities define the CDF of the continuous predictand at the category boundaries, and can be linearly interpolated to reconstruct the full predictive CDF. However, these variables at some grid points and times of the year may have a distribution with a combination of discrete and continuous characteristics. For example, it is possible that a grid cell could be cloud-free ($\text{tcc}_{\text{noon}} = 0.0$) or be completely cloud-covered ($\text{tcc}_{\text{noon}} = 1.0$), and the POLR method could predict a positive probability (i.e., point mass) for exactly these conditions. The CDF is discrete for these conditions and continuous for all other possible tcc_{noon} values ($0.0 < \text{tcc}_{\text{noon}} < 1.0$), meaning that the full CDF has a discontinuity at $(0.0, 0.0)$ and $(1.0, 1.0)$. We therefore use linear interpolation between the upper boundary of the first and lower boundary of the last category, and extrapolate beyond those boundaries with a constant value, so that the probability for the first (last) category becomes a point mass for the tcc_{noon} value of 0.0 (1.0). Likewise, we allow for a point mass at $(0, 0)$ and $(1, 1)$ for $\text{dp}_{24\text{h}}$, but we only allow for a point mass at $(1, 1)$ for RH, because 0% RH is unrealistic at the surface.

The inverse of the interpolated CDF (CDF^{-1}) allows us to generate forecasts of the continuous predictand based on the predicted category probabilities from the POLR method. This inverse CDF can then be sampled as described in section 5c to construct a finite ensemble of forecasts, which are needed as inputs to the GEFF.

4) SNOW COVER

The remaining meteorological input to the GEFF is the binary snow cover mask $\text{sc}_{\{m, \text{noon}\}}$ based on snow depth (sd) at local noon. The GEFF defines the snow mask as

$$\text{sc}_{\{m, \text{noon}\}} = \{1, \text{ if } \text{sd} \geq 1.0 \text{ mm}; 0, \text{ if } \text{sd} < 1.0 \text{ mm}\}. \quad (3)$$

Unlike the meteorological variables that used the POLR method, there are only two possible categories (snow or no snow). A common way to relate such a binary predictand to a continuous predictor like sd is to provide a probability of the binary outcome via logistic regression (Hamill et al. 2004; Wilks 2011). We use ensemble statistics $\mathbf{x} = (1, r_{\text{ENS}}, f_0)^T$ of sd as the predictors; while there is potential value in knowing how many ensemble members predict a snow depth of (near) zero, there is no need for predictor f_1 since there is no natural upper bound for snow depth that would motivate the definition of such a predictor. The training and verification period are the same that were used for temperature and wind speed. Technical details about the logistic-regression approach and how we generated an ensemble of binary postprocessed

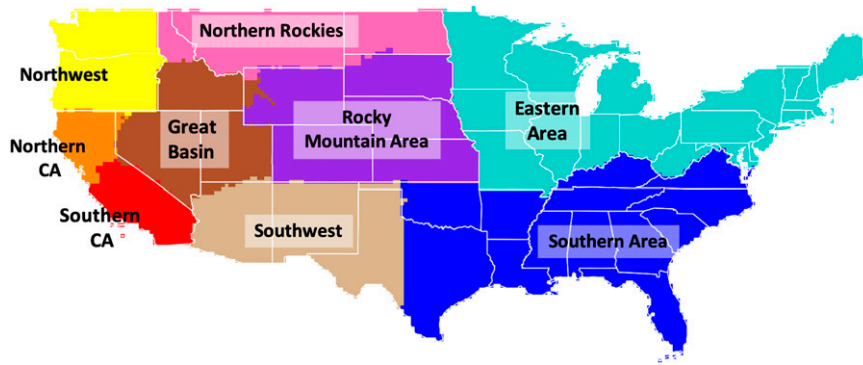


FIG. 2. The nine GACC regions within the contiguous United States.

forecasts $sc_{\{m, \text{noon}\}}$ from this method are in the supplemental material.

c. Ensemble copula coupling

All of the univariate postprocessing methods discussed in the previous subsections produce calibrated predictive distributions for each variable, lead time, and location. To generate a finite ensemble of postprocessed forecasts from these distributions to input to the GEFF system, we systematically sample the distributions at $M = K = 11$ equidistant quantile levels defined as $\tau_k = (k - 0.5)/k$, $k = 1, \dots, K$. These particular levels entail a CRPS-optimal sample (Bröcker 2012).

Because postprocessing was performed independently for each variable, lead time, and location, relationships between physically dependent variables such as precipitation and cloud cover are not considered. However, the intervariable, temporal, and spatial relationships should be restored so that the meteorological variables input to the equations in the GEFF system will lead to physically realistic fire-indicator forecasts that are spatially and temporally coherent. To restore the dependence structures, we employ a reordering method called ensemble copula coupling (ECC; Schefzik et al. 2013). The ECC method reorders the M postprocessed ensemble members according to the rank structure of the raw ensemble under the assumption that the raw ensemble is representative of observed intervariable, temporal, and spatial dependencies. Effectively, this approach will help instill any flow structures or dependencies back into the postprocessed forecasts including large-change events such as up-ramps in wind speed and drops in temperature during a cold frontal passage. The final result is an ensemble of forecasts with intervariable coherence as well as spatial and temporal coherence for each of the meteorological variables input to the GEFF system.

6. Verification results

a. Skill scores of a fire indicator at Geographic Area Coordination Centers

Logistical coordination and mobilization of fire mitigation and suppression resources at the federal and state levels within

the United States are conducted within 11 Geographic Area Coordination Centers (GACCs), 9 of which are within the CONUS (Fig. 2). These GACCs are under the direction of the National Interagency Fire Center (NIFC) and often communicate and share resources across GACC boundaries to enhance wildfire preparedness levels. Forecasters within these GACCs also monitor weather and fuels conditions to provide fire danger outlooks. Therefore, we show how the raw and postprocessed forecasts perform in each of these regions.

We quantify forecast performance using the skill of the continuous ranked probability score (CRPS). The CRPS summarizes the sharpness and calibration of the ensemble forecast (Gneiting et al. 2005; Gneiting and Raftery 2007) and its skill score (CRPSS) quantifies the performance relative to benchmark forecasts.

Climatological forecasts are natural benchmarks for extended-range forecasts, but we also use the raw forecasts as benchmarks to quantify the improvement gained from statistical postprocessing. The climatological benchmark forecasts were calculated from the ERA5 reanalysis data using a ± 5 -day sliding window around each day of the year for all available years from 1998 to 2018. Eleven equidistant quantiles were sampled from this climatological distribution (same as in section 5c) to match the size of the raw and postprocessed ensembles.

The BI is one of the terminus outputs from the GEFF and is often used to indicate how intense a fire would be and how swiftly it would spread in worst-case scenarios. The BI forecasts generated with postprocessed forecasts ran through the GEFF show improved skill relative to those generated with raw forecasts in all GACC regions and for all lead times out to day 14 (Fig. 3a) during the summer months (JJA). In general, the greatest improvement is seen in the far western GACC regions (warm colors in Fig. 3), and the Southern Area GACC.

Before passing the meteorological variables to the GEFF system, we verified that the postprocessing methods for each variable were beneficial and led to increased or at least as good of skill as the raw forecasts (in the form of positive CRPSS values). For this reason, the fire indicators calculated by the GEFF are inherently postprocessed with respect to ERA5 reanalyses as well. The GEFF outputs can draw on the individual and collective improvements of each meteorological variable gained from postprocessing. For example, the BI is

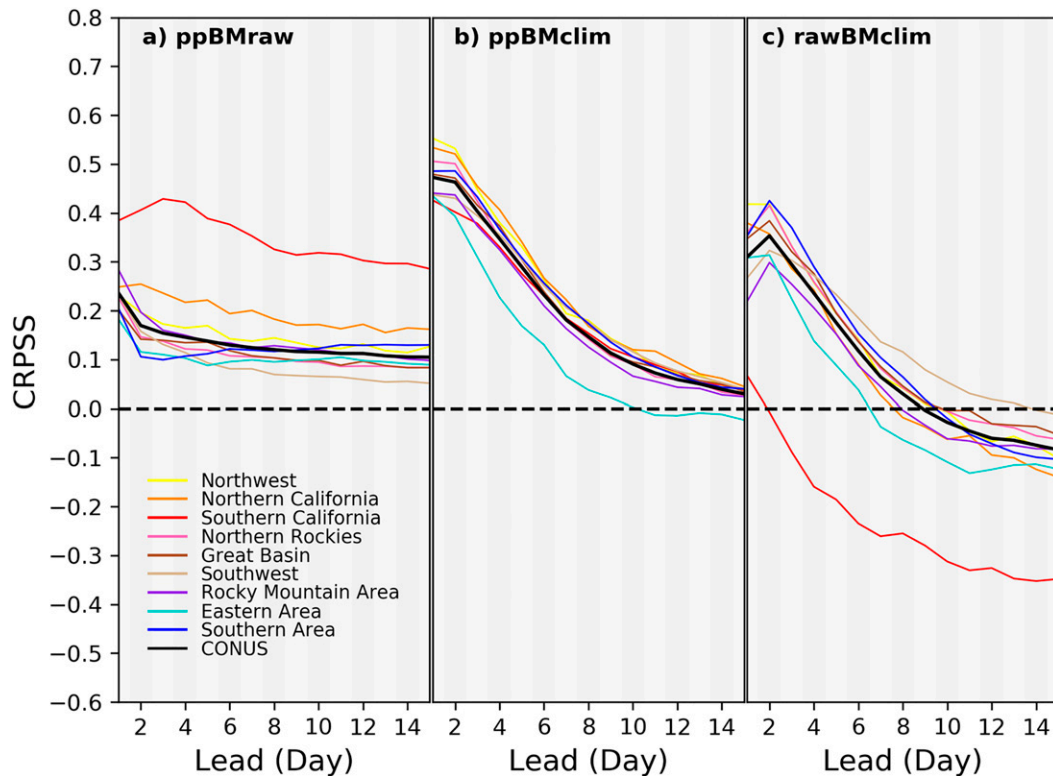


FIG. 3. Mean CRPSS values for the burning index (BI) calculated for each GACC region (solid colored lines) and for the entire CONUS domain (solid black line) for forecasts out to 14 days ahead. Skill of the (a) postprocessed forecasts relative to raw forecasts, (b) postprocessed forecasts relative to climatological forecasts, and (c) raw forecasts relative to climatological forecasts are shown. Values above (below) zero (dashed black line) indicate improvement (degradation) relative to the benchmark forecasts. The GACC regions and colors correspond to the regions in Fig. 2.

influenced by several factors, but the meteorological variables that showed the greatest enhancement relative to raw forecasts through postprocessing were the relative humidity variables (Table 3). In contrast, the wind speed and cloud cover variables show less improvement (i.e., smaller but still positive CRPSS values) and are therefore expected to contribute relatively less to the overall skill improvements seen in the BI forecasts.

The skill of the postprocessed forecasts relative to climatological forecasts during JJA is mostly uniform across the GACC regions and varies strongly with lead time (Fig. 3b). As expected, the skill decreases with increasing lead time, yet still shows positive skill in all GACC regions out to 14 days, except for the Eastern Area which only shows positive skill out to 10 days. In contrast, the raw forecasts lose skill relative to climatological forecasts in most GACC regions by day 9 and even sooner in some GACCs (Fig. 3c).

While the exact magnitude of the CRPSS and the day at which the raw forecast becomes worse than the climatological forecast varies for each GACC region and for each season (not shown), these overall general patterns remain: 1) postprocessing is beneficial over simply using the raw forecasts, 2) the raw forecasts are often less skillful than the climatological forecasts for days 8–14 while 3) the postprocessed forecasts are often more skillful than the climatological

forecasts even out to 14 days ahead. These results indicate that postprocessing is necessary for the most skillful fire-indicator forecasts at days 8–14. Tables of CONUS-wide mean CRPSS values for a subset of meteorological variables, fire indicators, and lead times during each season are available in the supplemental material.

b. Reliability of forecasts for above- and below-normal events

Extended range forecasts (D8–14) are often used to assess whether forecasted conditions will be above- or below-normal relative to local climatology. The CPC uses the lower (upper) tercile of a climatological distribution to define below- (above-) normal events for their week-two forecasts. We use this definition too and construct reliability diagrams to assess the performance of the raw and postprocessed forecasts at predicting these events with varying degrees of certainty.

To construct our local climatological distribution for each day of the year, we use 20 years of ERA5 data at a particular grid point with ± 5 days around each day of the year. To create a reliability diagram (as in Fig. 4) for each season, we aggregate these daily climatological distributions (20 years \times 11 days) for all days of the year for which there is a corresponding ECMWF reforecast within a 3-month period (e.g., June–July–August).

TABLE 3. CONUS-wide mean CRPSS values for a subset of meteorological and fire-indicator values input and output from the GEFF, respectively. The fire indicators were generated with inputs from the raw and postprocessed ECMWF ensemble reforecasts. The CRPSS values were calculated with 20 years of reforecasts in JJA and use the raw forecasts and climatological forecasts (± 7 days around date, 20 years) as benchmarks for comparison. Values above (below) zero indicate improvement (degradation) relative to the benchmark; ppBMraw = postprocessed GEFF forecasts relative to raw GEFF forecasts; ppBMclim = postprocessed GEFF forecasts relative to climatological GEFF forecasts; rawBMclim = raw GEFF forecasts relative to climatological GEFF forecasts.

Mean CRPSS over CONUS domain: Jun–Jul–Aug						
Variable	Forecast vs benchmark	D6	D8	D10	D12	D14
tcc _{noon}	ppBMraw	0.05	0.058	0.061	0.062	0.063
	ppBMclim	0.101	0.046	0.016	0.005	−0.001
	rawBMclim	0.054	−0.014	−0.048	−0.061	−0.068
RH _[24h,min]	ppBMraw	0.297	0.267	0.249	0.235	0.231
	ppBMclim	0.251	0.163	0.105	0.068	0.051
	rawBMclim	−0.066	−0.142	−0.191	−0.219	−0.235
t _{noon}	ppBMraw	0.18	0.157	0.149	0.132	0.124
	ppBMclim	0.299	0.178	0.098	0.053	0.032
	rawBMclim	0.144	0.025	−0.06	−0.09	−0.105
tp _{24h}	ppBMraw	0.072	0.064	0.069	0.067	0.07
	ppBMclim	0.132	0.069	0.028	0.011	0.008
	rawBMclim	0.065	0.005	−0.044	−0.06	−0.066
u _{noon}	ppBMraw	0.081	0.077	0.073	0.076	0.073
	ppBMclim	0.123	0.048	0.019	0.002	0.003
	rawBMclim	0.046	−0.032	−0.059	−0.081	−0.076
bi	ppBMraw	0.13	0.12	0.115	0.112	0.105
	ppBMclim	0.232	0.147	0.091	0.059	0.039
	rawBMclim	0.118	0.03	−0.028	−0.06	−0.074
erc	ppBMraw	0.149	0.141	0.135	0.133	0.126
	ppBMclim	0.331	0.231	0.159	0.109	0.081
	rawBMclim	0.215	0.105	0.028	−0.028	−0.051
fwi	ppBMraw	0.188	0.162	0.152	0.145	0.133
	ppBMclim	0.314	0.21	0.144	0.096	0.075
	rawBMclim	0.156	0.056	−0.01	−0.058	−0.066
ic	ppBMraw	0.166	0.148	0.138	0.131	0.119
	ppBMclim	0.205	0.119	0.065	0.038	0.022
	rawBMclim	0.047	−0.034	−0.086	−0.107	−0.11

From this 3-month climatological distribution, we sample the terciles (i.e., the 33.33% and 66.66% percentiles) like CPC does to determine the threshold for below- and above-normal events.

The reliability diagrams in Fig. 4 show the full joint distributions between probabilistic raw day-8 forecasts and the observations (i.e., ERA5-based GEFF runs) for binary BI events during the summertime in each GACC region. Histograms of the count of each of the probability forecasts in the reliability diagrams are available in the supplemental material.

The raw forecasts show three main features: 1) they are overconfident for moderate to high forecast probabilities, 2) they suffer from poor resolution and therefore cannot discern forecasts for which the outcomes are substantially different from each other, and 3) the reliability varies for each GACC region. These features are more exaggerated for above-normal events. These results indicate that the average raw forecast does not exhibit large enough spread, which is a common deficiency of raw NWP ensembles. Additionally, the raw forecasts do not provide a sufficiently strong probabilistic signal for above- and below-normal conditions, especially for higher probabilities of above-normal events. For example, relative frequencies of a $\sim 60\%$ forecast probability is not all that different from those

with a $\sim 80\%$ probability; the observed relative frequencies depend only weakly on the probabilistic forecasts. Last, since the reliability varies for each GACC region, this would make it difficult for a forecaster to issue consistently reliable forecasts across the entire CONUS domain.

Postprocessing the meteorological forecasts before passing them through the GEFF system mitigates all of the deficiencies exhibited by the raw forecasts. The reliability diagrams (Fig. 5) show that the probabilistic forecasts are much closer to the 1:1 line meaning that the forecasts are better calibrated and are no longer overly confident. They show good resolution and the reliability is consistently improved for each GACC region (except for high probabilities of above normal BI events in the Eastern Area, which is likely a result of few high probability forecasts of extreme BI values in this typically wet region). The consistency in reliability among different GACC regions is a result of the gridpoint-by-gridpoint postprocessing techniques that we used. Reliability diagrams for D12 (not shown) depict slightly less reliability than D8, but overall show similar results to D8.

c. The impact of aggregating forecast days

As forecasts evolve further and further from the constraints of the initial conditions, the model struggles to predict instantaneous

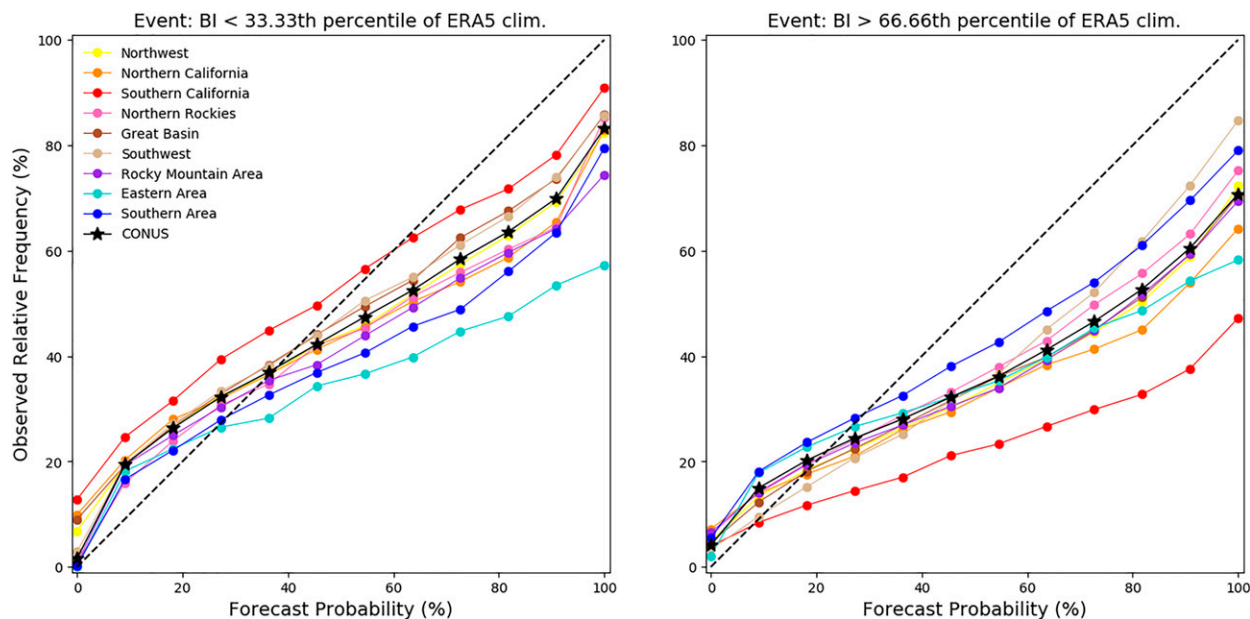


FIG. 4. Reliability diagrams for daily raw forecasts during the summertime (June–July–August) for burning index (BI) values (left) less than the ERA5 climatological lower tercile and (right) greater than the ERA5 climatological upper tercile in each GACC region (colored lines) and for the entire CONUS domain (black line with star symbols). Forecasts are for 8 days ahead and were constructed, along with the ERA climatology, for the period from 1998 to 2018. ERA5 data are used for observations. The $N = 12$ possible probabilities are equal to $p = (k \times 100)/(N - 1)$, $k = 0, 1, \dots, N - 1$. The dashed diagonal line represents perfect calibration.

states of the atmosphere (Epstein 1988) leading to timing errors in the forecast. One method to enhance skill at these longer horizons is to calculate aggregate statistics such as averages over a span of several days. The Climate Prediction Center currently creates aggregated temperature and precipitation forecast outlooks for days 8–14 so we show how this can be advantageous for fire-indicator forecasts too.

We generate a days 8–14 aggregated forecast for each re-forecast date and ensemble member by averaging post-processed forecast values for day 8 up to and including day 14. To evaluate the skill of these outlook forecasts compared to climatology, we sample 11 equidistant quantiles from an ERA5 aggregated climatological distribution at each location. The distribution includes averaged values valid for the

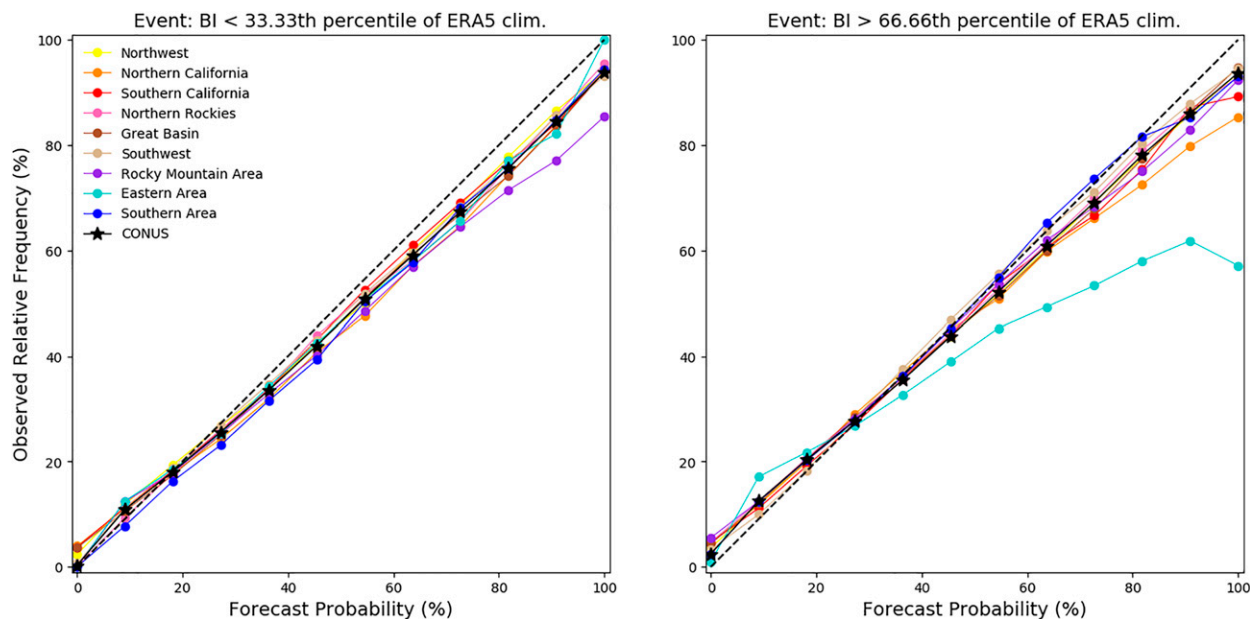


FIG. 5. As in Fig. 4, but for postprocessed forecasts.

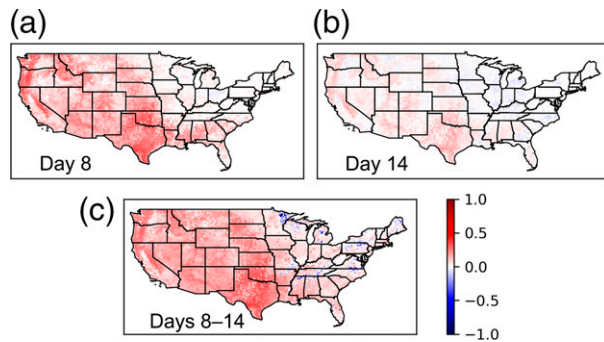


FIG. 6. CRPSS values of the BI during the summer for (a) day 8 and (b) day 14. (c) CRPSS values calculated with averaged postprocessed forecasts and averaged climatological reference forecasts spanning days 8–14. Red-shaded regions indicate positive skill over a reference climatological forecast.

same days 8–14 outlook period for each of the 20 available years. Using the aggregated reforecasts and reanalyses, we calculate CRPS values and subsequent CRPSS values using a benchmark of climatology for each season and lead time.

Aggregating postprocessed BI forecasts over days 8–14 still yields positive skill compared to a reference climatological forecast for the same period across the majority of the CONUS domain during the summer (Fig. 6). Additionally, the skill of the aggregated forecast is overall greater than forecasts for just a single day (e.g., day 8, day 12), even day 8 as shown by the deeper red shades in Fig. 6c than in Figs. 6a and 6b. This improvement is particularly noticeable when comparing the aggregated forecasts with a single day-14 forecast. Consider the following example to demonstrate the utility of this type of fire forecast. To redistribute shared resources across large areas, a stakeholder may not need to know which single day will have the worst fire conditions but just that the forecasted fire conditions show above-normal conditions during week two. Aggregating forecasts will enable the stakeholder to make a more confident decision than if they were to refer to any single day within the days 8–14 range.

7. Conclusions

We developed a framework that combines several statistical postprocessing methods and the GEFF system to produce skillful probabilistic fire-indicator forecasts for days 8–14. We implemented four different parametric-based postprocessing methods that were suitable for the eleven different meteorological variables. The variables included those that most influence fire behavior, spread, and intensity: cloud and snow cover, surface temperature, relative humidity, and wind speed and accumulated precipitation quantities. Our postprocessing methods benefited from a large, 20-yr dataset of ECMWF reforecasts and corresponding ERA5 reanalysis data for training and verification. These reforecasts were key for the postprocessing models to identify and correct for systematic biases in the raw ensemble forecasts, especially for extended lead times. While we directly ran the GEFF with ERA5 reanalysis data for verification, ECMWF has made a GEFF reanalysis

dataset based on ERA5 (Vitolo et al. 2020) available to the public at: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-fire-historical?tab=overview>.

We found that postprocessing the meteorological forecasts input to the GEFF system led to improved skill of the fire-indicator forecasts relative to benchmarks of raw and climatological forecasts. We showed results for the burning index (BI) fire indicator during the summertime, but similar results were found for the other NFDRS fire-indicators and seasons. The main takeaways are that statistical postprocessing yielded more skillful BI forecasts relative to raw forecasts in all GACC regions and for all lead times out to day 14. We found the postprocessed forecasts relative to climatological forecasts to have positive skill in all GACC regions. However, the skill degraded gradually with increased lead time. Even still, we can expect positive skill in days 8–14 relative to a climatological forecast. In contrast, if the raw forecast was used, the skill became negative around days 8/9 and even sooner in some GACC regions. These results underscore the importance of statistical postprocessing to not only improve the skill of the forecasts, but to even get any usable skill from NWP fire forecasts at days 8–14. We quantified the reliability of above- and below-normal forecasts of the BI and found that the raw forecasts were overconfident and the degree of reliability varied for each GACC region. Conversely, the postprocessed forecasts were no longer overconfident and were consistently reliable across all GACC regions, making it easier for a forecaster to issue a more consistent nationwide forecast. Last, we examined the effect of aggregating the postprocessed forecast days on the skill scores. Averaging the forecasts over the days 8–14 period led to larger positive skill scores across the CONUS domain than those generated for any single day in this period, including day 8. This result suggests that the most skillful fire-indicator forecasts for week two are aggregated ones.

We trained the postprocessing models using ERA5 reanalyses as a proxy for “observed” outcomes. Additionally, we calculated the verification metrics using ERA5-based GEFF outputs (i.e., ERA5 data ran through the GEFF system) rather than observed fire events. While verification with meteorological data from observation sites and actual wildfire events could be useful to train postprocessing models, it presents many challenges. The observed wildfires do not allow for a gridpoint-by-gridpoint verification and even if all of the weather and fuels conditions are primed for a fire, a fire will not occur unless there is an ignition, which leads to false alarms. Therefore, the ERA5 reanalyses allowed for the best assessment of the impact of our statistical postprocessing methods on fire-indicator forecasts.

The framework and postprocessing algorithms developed for this paper are being repeated with the latest GEFSv12 reforecasts (NOAA 2020) and then will be transferred to the Climate Prediction Center for use in near-real-time operations. The forecasts, in the form of probabilistic above- and below-normal maps, will complement shorter-lead fire-weather and fire-indicator forecasts offered by the NWS and the Wildland Fire Assessment System. This paper focused on BI forecasts, because of limited space and because the BI is commonly used by fire managers to make strategic decisions. Since the GEFF system outputs fuel responses and fire-indicator forecasts from

three different fire danger ratings systems, any of the components of those rating systems could be used as forecast tools. This study was limited to the CONUS domain, but the postprocessing methods in combination with the GEFF system have the potential to produce skillful week-two forecasts anywhere if given the representative meteorological and fuels data for the region. Future work will try to leverage machine-learning capabilities to directly predict fire indicators and streamline the process in operations by bypassing many of the individual postprocessing steps.

Acknowledgments. We are grateful for the many researchers, forecasters, and end-users from the USDA Forest Service and the National Weather Service offices and prediction centers who offered suggestions and feedback throughout this project. We are thankful for access to ECMWF's operational medium-range forecasts and the GEFF-system code provided by Francesca Di Giuseppe and Xavier Abellan. We also greatly appreciate the insightful and constructive feedback from three anonymous reviewers, which led to improvements of this paper. This research was supported by a grant from a Service Level Agreement between NOAA/ESRL/PSL and NOAA/NWS/CPC to develop fire-weather related 8–14-day forecast products at the CPC and by the NOAA Cooperative Agreement with CIRES (Grant NA17OAR4320101).

Data availability statement. The reforecasts used in this study were downloaded from the ECMWF Meteorological Archival and Retrieval System (MARS; <https://www.ecmwf.int/en/forecasts/access-forecasts/access-archive-datasets>) with special permission from the ECMWF. The ERA5 reanalysis data are publicly available from the Copernicus Climate Change Service Climate Data Store (C3S 2018). Open-source code to run the GEFF system is available at <https://github.com/ecmwf-projects/geff> and precomputed GEFF runs based on ERA5 reanalyses are now available through the C3S at: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-fire-historical?tab=overview>.

REFERENCES

- Bedia, J., N. Golding, A. Casanueva, M. Iturbide, C. Buontempo, and J. M. Gutiérrez, 2018: Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. *Climate Serv.*, **9**, 101–110, <https://doi.org/10.1016/j.cliser.2017.04.001>.
- Bowman, D. M. J. S., and Coauthors, 2009: Fire in the Earth system. *Science*, **324**, 481–484, <https://doi.org/10.1126/science.1163886>.
- Bradshaw, L. S., J. E. Deenning, R. E. Burgan, and J. D. Cohen, 1983: The 1978 National Fire-Danger Rating System: Technical documentation. Tech. Rep. INT-169, USDA Forest Service, 49 pp., https://www.fs.fed.us/rm/pubs_int/int_gtr169.pdf.
- Bröcker, J., 2012: Evaluating raw ensembles with the continuous ranked probability score. *Quart. J. Roy. Meteor. Soc.*, **138**, 1611–1617, <https://doi.org/10.1002/qj.1891>.
- Camia, A., P. Barbosa, G. Amatulli, and J. San-Miguel-Ayán, 2006: Fire danger rating in the European Forest Fire Information System (EFFIS): Current developments. *For. Ecol. Manage.*, **234**, S20, <https://doi.org/10.1016/j.foreco.2006.08.036>.
- C3S, 2018: ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service Climate Data Store, accessed 5 May 2021, <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview>.
- Cheng, L., M. Hoerling, Z. Liu, and J. Eischeid, 2019: Physical understanding of human-induced changes in U.S. hot droughts using equilibrium climate simulations. *J. Climate*, **32**, 4431–4443, <https://doi.org/10.1175/JCLI-D-18-0611.1>.
- Cohen, J., and J. Deeming, 1985: The National Fire Danger Rating System: Basic equations. Tech. Rep. PSW-82, USDA Forest Service, Pacific Southwest Forest and Range Experiment Station, 23 pp., https://www.fs.fed.us/psw/publications/documents/psw_gtr082/psw_gtr082.pdf.
- Di Giuseppe, F., F. Pappenberger, F. Wetterhall, B. Krzeminski, A. Camia, G. Libertá, and J. San Miguel, 2016: The potential predictability of fire danger provided by numerical weather prediction. *J. Appl. Meteor. Climatol.*, **55**, 2469–2491, <https://doi.org/10.1175/JAMC-D-15-0297.1>.
- , C. Vitolo, B. Krzeminski, C. Barnard, P. Maciel, and J. San-Miguel, 2020: Fire Weather Index: The skill provided by the European Centre for Medium-Range Weather Forecasts ensemble prediction system. *Nat. Hazards Earth Syst. Sci.*, **20**, 2365–2378, <https://doi.org/10.5194/nhess-20-2365-2020>.
- ECMWF, 2018: Summary of cycle 45r1. ECMWF, accessed 5 April 2021, <https://www.ecmwf.int/en/forecasts/documentation-and-support/evolution-ifs/cycles/summary-cycle-45r1>.
- Epstein, E. S., 1988: Long-range weather prediction: Limits of predictability and beyond. *Wea. Forecasting*, **3**, 69–75, [https://doi.org/10.1175/1520-0434\(1988\)003<0069:LRWPLO>2.0.CO;2](https://doi.org/10.1175/1520-0434(1988)003<0069:LRWPLO>2.0.CO;2).
- Flannigan, M., A. S. Cantin, W. J. de Groot, M. Wotton, A. Newbery, and L. M. Gowman, 2013: Global wildland fire season severity in the 21st century. *For. Ecol. Manage.*, **294**, 54–61, <https://doi.org/10.1016/j.foreco.2012.10.022>.
- , B. J. Stocks, and B. M. Wotton, 2000: Climate change and forest fires. *Sci. Total Environ.*, **262**, 221–229, [https://doi.org/10.1016/S0048-9697\(00\)00524-6](https://doi.org/10.1016/S0048-9697(00)00524-6).
- Fu, R., A. Howell, J. Mankin, A. Sheffield, and I. Simpson, 2021: Tackling challenges of a drier, hotter, more fire-prone future. *Eos*, accessed 4 April 2021, <https://eos.org/opinions/tackling-challenges-of-a-drier-hotter-more-fire-prone-future>.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , —, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, [https://doi.org/10.1175/1520-0493\(2004\)132<1434:ERIMFS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2).
- Hemri, S., T. Haiden, and F. Pappenberger, 2016: Discrete post-processing of total cloud cover ensemble forecasts. *Mon. Wea. Rev.*, **144**, 2565–2577, <https://doi.org/10.1175/MWR-D-15-0426.1>.
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- , and Coauthors, 2018: ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 5 April 2021, <https://doi.org/10.24381/cds.adbb2d47>.
- Jolly, W. M., M. A. Cochrane, P. H. Freeborn, Z. A. Holden, T. J. Brown, G. J. Williamson, and D. M. J. S. Bowman, 2015: Climate-

- induced variations in global wildfire danger from 1979 to 2013. *Nat. Commun.*, **6**, 7537, <https://doi.org/10.1038/ncomms8537>.
- McArthur, A. G., 1966: *Weather and Grassland Fire Behaviours*. Australia Forestry and Timber Bureau, 23 pp., <https://catalogue.nla.gov.au/Record/752731>.
- Mölders, N., 2010: Comparison of Canadian Forest Fire Danger Rating System and National Fire Danger Rating System fire indices derived from Weather Research and Forecasting (WRF) model data for the June 2005 Interior Alaska wildfires. *Atmos. Res.*, **95**, 290–306, <https://doi.org/10.1016/j.atmosres.2009.03.010>.
- Nabizadeh, E., P. Hassanzadeh, D. Yang, and E. A. Barnes, 2019: Size of the atmospheric blocking events: Scaling law and response to climate change. *Geophys. Res. Lett.*, **46**, 13 488–13 499, <https://doi.org/10.1029/2019GL084863>.
- NOAA, 2020: NOAA's Global ensemble forecast system version 12: Reforecast data storage information. NOAA, 7 pp., https://noaa-gefs-retrospective.s3.amazonaws.com/Description_of_reforecast_data.pdf.
- Parks, S. A., C. Miller, M.-A. Parisien, L. M. Holsinger, S. Z. Dobrowski, and J. Abatzoglou, 2015: Wildland fire deficit and surplus in the western United States, 1984–2012. *Ecosphere*, **6**, 1–13, <https://doi.org/10.1890/ES15-00294.1>.
- Radeloff, V. C., and Coauthors, 2018: Rapid growth of the US wildland-urban interface raises wildfire risk. *Proc. Natl. Acad. Sci. USA*, **115**, 3314–3319, <https://doi.org/10.1073/pnas.1718850115>.
- Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, <https://doi.org/10.1214/13-STS443>.
- Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- , and —, 2018: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *J. Hydrometeorol.*, **19**, 1651–1670, <https://doi.org/10.1175/JHM-D-18-0067.1>.
- Schlobohm, P., and Coauthors, 2002: Gaining an understanding of the National Fire Danger Rating System. USDA National Wildfire Coordinating Group, Rep. PMS 932/NFES 2665, 82 pp., <https://www.nwcg.gov/sites/default/files/products/pms932.pdf>.
- Schoennagel, T., and Coauthors, 2017: Adapt to more wildfire in western North American forests as climate changes. *Proc. Natl. Acad. Sci. USA*, **114**, 4582–4590, <https://doi.org/10.1073/pnas.1617464114>.
- SPC, 2021: SPC Fire Weather Forecast Criteria. NOAA, 7 pp., https://www.spc.noaa.gov/misc/Critical_Criteria_for_web.pdf.
- Srock, A. F., J. J. Charney, B. E. Potter, and S. L. Goodrick, 2018: The hot-dry-windy index: A new fire weather index. *Atmosphere*, **9**, 279, <https://doi.org/10.3390/atmos9070279>.
- Stauffer, R., N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, <https://doi.org/10.1175/MWR-D-16-0260.1>.
- Van Wagner, C. E., 1974: Structure of the Canadian Forest Fire Weather Index. Canadian Forestry Service, 49 pp., <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.460.3231&rep=rep1&type=pdf>.
- , 1987: Development and structure of the Canadian Forest Fire Weather Index System. Canadian Forestry Service, 48 pp., <https://d1ied5g1xfp8x8.cloudfront.net/pdfs/19927.pdf>.
- Vitolo, C., F. Di Giuseppe, C. Barnard, R. Coughlan, J. San-Miguel-Ayanz, G. Libertá, and B. Krzeminski, 2020: ERA5-based global meteorological wildfire danger maps. *Sci. Data*, **7**, 216, <https://doi.org/10.1038/s41597-020-0554-z>.
- Wang, D., and Coauthors, 2021: Economic footprint of California wildfires in 2018. *Nat. Sustain.*, **4**, 252–260, <https://doi.org/10.1038/s41893-020-00646-7>.
- Westerling, A. L., H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam, 2006: Warming and earlier spring increase western U.S. forest wildfire activity. *Science*, **313**, 940–943, <https://doi.org/10.1126/science.1128834>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- Williams, A. P., J. T. Abatzoglou, A. Gershunov, J. Guzman-Morales, D. A. Bishop, J. K. Balch, and D. P. Lettenmaier, 2019: Observed impacts of anthropogenic climate change on wildfire in California. *Earth's Future*, **7**, 892–910, <https://doi.org/10.1029/2019EF001210>.
- Worsnop, R. P., M. Scheuerer, and T. M. Hamill, 2020: Extended-range probabilistic fire-weather forecasting based on ensemble model output statistics and ensemble copula coupling. *Mon. Wea. Rev.*, **148**, 499–521, <https://doi.org/10.1175/MWR-D-19-0217.1>.
- Xu, R., and Coauthors, 2020: Wildfires, global climate change, and human health. *N. Engl. J. Med.*, **383**, 2173–2181, <https://doi.org/10.1056/NEJMSr2028985>.