# High-Resolution Rapid Refresh Model Data Analytics Derived on the Open Science Grid to Assist Wildland Fire Weather Assessment

BRIAN K. BLAYLOCK

*Department of Atmospheric Sciences, University of Utah, Salt Lake City, Utah*

JOHN D. HOREL AND CHRIS GALLI

*Department of Atmospheric Sciences, University of Utah, and Synoptic Data, Salt Lake City, Utah*

(Manuscript received 9 May 2018, in final form 3 October 2018)

ABSTRACT

Terabytes of weather data are generated every day by gridded model simulations and in situ and remotely sensed observations. With this accelerating accumulation of weather data, efficient computational solutions are needed to process, archive, and analyze the massive datasets. The Open Science Grid (OSG) is a consortium of computer resources around the United States that makes idle computer resources available for use by researchers in diverse scientific disciplines. The OSG is appropriate for high-throughput computing, that is, many parallel computational tasks. This work demonstrates how the OSG has been used to compute a large set of empirical cumulative distributions from hourly gridded analyses of the High-Resolution Rapid Refresh (HRRR) model run operationally by the Environmental Modeling Center of the National Centers for Environmental Prediction. These cumulative distributions derived from a 3-yr HRRR archive are computed for seven variables, over 1.9 million grid points, and each hour of the calendar year. The HRRR cumulative distributions are used to evaluate near-surface wind, temperature, and humidity conditions during two wildland fire episodes—the North Bay fires, a wildfire complex in Northern California during October 2017 that was the deadliest and costliest in California history, and the western Oklahoma wildfires during April 2018. The approach used here illustrates ways to discriminate between typical and atypical atmospheric conditions forecasted by the HRRR model. Such information may be useful for model developers and operational forecasters assigned to provide weather support for fire management personnel.

## 1. Introduction

Weather data are generated at an enormous rate, so much that novel compute strategies are required to handle the velocity, volume, value, variety, and veracity of such ''big data.'' Massive datasets are produced by ensemble forecast systems (Schwartz et al. 2015; Loeser et al. 2017), high-spatial- and high-temporal-resolution numerical models (Benjamin et al. 2016), advanced next-generation weather satellites (Schmit et al. 2017), Internet of Things devices (Chapman and Bell 2018), and a growing number of other observational and model sources. The continued and accelerated production of data by these sources and forthcoming technologies poses data management and data analysis challenges common to other disciplines. The National Oceanic and Atmospheric Administration (NOAA) Big Data Project is testing the feasibility of storing these datasets on the public cloud (Ansari et al. 2018). Blaylock et al. (2017a) used similar storage technologies and created a publicly accessible archive of the High-Resolution Rapid Refresh (HRRR) model output. They demonstrated different use cases for using retrospective HRRR analyses and forecasts, which included using HRRR model output as initial and boundary conditions for the Weather Research and Forecasting Model (Blaylock et al. 2017b). Over 280 users have downloaded files from this HRRR archive since its inception.

Even when massive amounts of model and observational data are available and made accessible, it can be difficult to analyze entire datasets using a limited number of compute nodes. Furthermore, it is inefficient to use high-performance computing center resources to undertake repetitive computational tasks. Instead, the

*Corresponding author*: Brian K. Blaylock, brian.blaylock@utah.edu

DOI: 10.1175/JTECH-D-18-0073.1

dimensions of datasets are often reduced by focusing on case studies, considering only a restricted time window or region for analysis, or averaging over entire days or months. In addition, data mining techniques applied to a handful of variables are common along with the increased use of artificial intelligence to support decision-making during high-impact weather events (McGovern et al. 2017). However, to take advantage of the information available, it may not be appropriate to assume in advance how the dimensionality of the dataset should be reduced.

Scientists in many disciplines use grid and cloud computing infrastructures to manage workflows for processing large datasets (Juve et al. 2013). The Open Science Grid (OSG; https://www.opensciencegrid.org/) is a distributed high-throughput computing (HTC) system capable of allocating many computer resources from over 100 sites (Foster 2005; Pordes et al. 2007, 2008). HTC should not be confused with high-performance computing (HPC) systems, where highly optimized codes, often requiring extensive message passing between computer nodes, are intended to be completed in the shortest amount of time possible. In contrast, HTC systems are designed to complete many smaller jobs undertaken independently of each other, which is a useful approach to mine massive datasets.

The objective of this work was to compute empirical cumulative distributions from a multiyear archive of HRRR output for several near-surface variables relevant to fire weather applications. Such statistics may help fire managers and incident meteorologists understand model behavior and weather conditions in the vicinity of wildland fires. As summarized by Page et al. (2018), having confidence in the accuracy of weather forecast guidance is necessary for wildland fire professionals to make critical strategic decisions and inform tactical operations. National Weather Service (NWS) forecasters at Weather Forecast Offices and incident meteorologists on location at major fires depend on observations and model forecast guidance for situational awareness in locations for which they may have limited prior experience (Gravelle et al. 2016; Powers et al. 2017). Forecasters need to be able to recognize what are typical and atypical atmospheric conditions at that time of year and the extent to which the available model guidance successfully forecasts those conditions. A comparison of observed to modeled cumulative distributions also helps identify model biases over the range of conditions observed (Wilks 2011).

The OSG computing resources available to researchers are described in section 2. An overview of the HRRR model and statistical methodology applied to both model output and surface observations are given in section 3. Section 4 illustrates how the empirical cumulative distributions calculated from the HRRR F00 fields may be used to identify atypical atmospheric conditions for any location in the HRRR domain throughout the year. Cumulative distributions of HRRR analyzed 10-m wind speed are compared to HRRR forecasted winds for the onset period of the northern California North Bay fires during 8–9 October 2017 and the western Oklahoma fires on 13 April 2018. We also compare the cumulative distributions of the HRRR winds to those observed at nearby stations that generally have much longer records, with particular attention placed on the 2000–present record at the Hawkeye (HWKC1) Remote Automatic Weather Station (RAWS) in the vicinity of one of the North Bay fires.

## 2. The OSG

The OSG is a worldwide consortium of distributed computing resources (Pordes et al. 2007, 2008; Sfiligoi et al. 2009). Members of the consortium make their idle computing resources available to other researchers, many of whom would otherwise not have access to such facilities (Pordes et al. 2007). OSG has been used for scientific workflows in particle physics (Norman et al. 2015; Chekanov et al. 2017), astronomy (Juve et al. 2013; Huerta et al. 2017), chemistry, genetics (Poehlman et al. 2016), and other science disciplines but sparingly in the atmospheric science community (e.g., Mülmenstädt et al. 2012; Houston et al. 2015). A wide variety of software packages are available and ready to use on the OSG. For specific and unique computational tasks that require specialized or custom software, users may run software packages using Docker and Singularity containers (https://www.docker.com/; https://www.sylabs.io/), which are common methods to virtualize software to run on diverse computer resources. Considerable documentation about the OSG is available online (https://support.opensciencegrid.org).

The OSG provides the infrastructure to test, complete, and recompute data mining and statistical computations applied to large amounts of input data in less time than traditional computing workflows. The OSG is most appropriate for performing computational tasks that are embarrassingly parallel—a term used in computational applications to refer to problems that can be decomposed easily into repetitive and independent tasks (Stockinger et al. 2006; Régin et al. 2013; Steward et al. 2017). While the OSG makes many compute resources available, queued tasks are completed independently of each other, since there is no shared file system between the workers. Using the OSG is not appropriate for some computational applications, such as those that rely on message passing interface required to run numerical weather prediction models, for example,

the Weather Research and Forecasting Model (Powers et al. 2017). However, using the OSG is a good resource to analyze the large datasets typically obtained from regional or global modeling systems. HTC workloads typically focus on acting on a high volume of parallel input rather than optimizing for time-sensitive computations.

Researchers can submit thousands of jobs from the submit node, OSG Connect (https://osgconnect.net/), to a queue managed by HTCondor (http://research.cs.wisc.edu/htcondor/). Although hardware and software among available worker nodes are heterogeneous, users can specify machine requirements for their jobs. For instance, upon job submission, users may request a certain operating system, a minimum amount of memory, and a minimum amount of disk space; and require the ability to run in a Singularity container. HTCondor then matches each job with resources in the OSG network that satisfy the requested requirements and sends the input data and any required software to worker nodes for execution when those resources become available. Jobs can be further managed and controlled with HTCondor's Directed Acyclic Graph Manager (DAGMan), which is necessary if one or more computational tasks depend on a previous task being completed. After a job finishes, any output is sent back to the user's home directory on the OSG submit node and can be stored on the OSG "Stash" archive, where users have a higher storage limit. If the user wishes to do further analysis on the generated output, that data may be transferred to her/his institution's home computing facilities or a personal computer with simple secure copy protocol. Bulk file transfers can be accomplished with Globus (https://www.globus.org/), a file transfer solution. Since submitted jobs use opportunistic cycles, they are at risk of being preempted by the compute resource owner. Hence, the OSG cannot be relied on for time-sensitive operational jobs. However, there are built-in provisions to restart preempted jobs so that all are completed.

## 3. Data and methods

### a. HRRR model

The HRRR forecast modeling system developed by NOAA's Earth System Research Laboratory and run operationally by the Environmental Modeling Center (EMC) produces F00–F18 forecasts for the contiguous United States at 3-km grid spacing (over 1.9 million grid points). An advanced data assimilation cycle is run to assimilate available observations every hour, with 15-min radar reflectivity used to specify a three-dimensional latent heating structure during a 1-h "preforecast" prior to the F00 model analysis (Hwang et al. 2015; Benjamin et al. 2016; James and Benjamin 2017). HRRR model output is used for nowcasting, situational awareness, and short-term forecasts for diverse operational applications, including aviation, solar and wind energy, agriculture, severe weather forecasting, and wildland fire management.

As described by Blaylock et al. (2017a), HRRR model datasets are retained and accessible for 48 h from servers at the National Centers for Environmental Prediction and not presently archived by the National Centers for Environmental Information. We began archiving operational HRRR output in April 2015 on the Pando archive system at the University of Utah's Center for High Performance Computing. Pando is a highly efficient object-storage archive system similar to Amazon Web Services S3 cloud storage. Each file on Pando has a unique Hypertext Transfer Protocol Secure (HTTPS) URL and can be accessed with a simple web get (wget) or client URL (cURL) command. Storing HRRR output in this manner gives remote servers within the OSG network easy access to download thousands of HRRR files in the archive for statistical processing.

### b. Statistical calculations on the OSG

Empirical cumulative distributions of observational and model datasets are commonly generated for specific locations and variables for a variety of applications (Wilks 2011; Hoffman et al. 2017; Zhao et al. 2017; Oswald 2018). For example, many operational forecasters relate current upper-air conditions to specific percentiles computed by the Storm Prediction Center from rawinsonde climatologies as a function of time of year (Rogers et al. 2014; http://www.spc.noaa.gov/exper/soundingclimo/). Multiyear model statistics from regional and global operational and research models are also ubiquitous but are generally limited to a few variables, and the statistics are often generalized to mean and standard deviation values or limited quantile sets over an entire year or season. For example, James et al. (2017, 2018) used a 3-yr dataset of F01 HRRR forecasts to estimate mean wind and solar energy resources within the HRRR domain.

We determined empirical cumulative distributions at each of the 1.9 million model grid points for F00 forecasts for all 8784 h of the calendar year and seven model output variables representative of those likely to be of interest for fire weather situational awareness: 2-m temperature, 2-m dewpoint temperature, 10-m instantaneous wind speed and gust, 10-m hour maximum wind speed, 80-m instantaneous wind speed, simulated composite reflectivity, and 500-hPa geopotential height. We limit our evaluation in this work to F00 fields only, since our archive of F01–F18 forecast fields begins in July 2016. To obtain a reasonably sized sample, the statistics for each

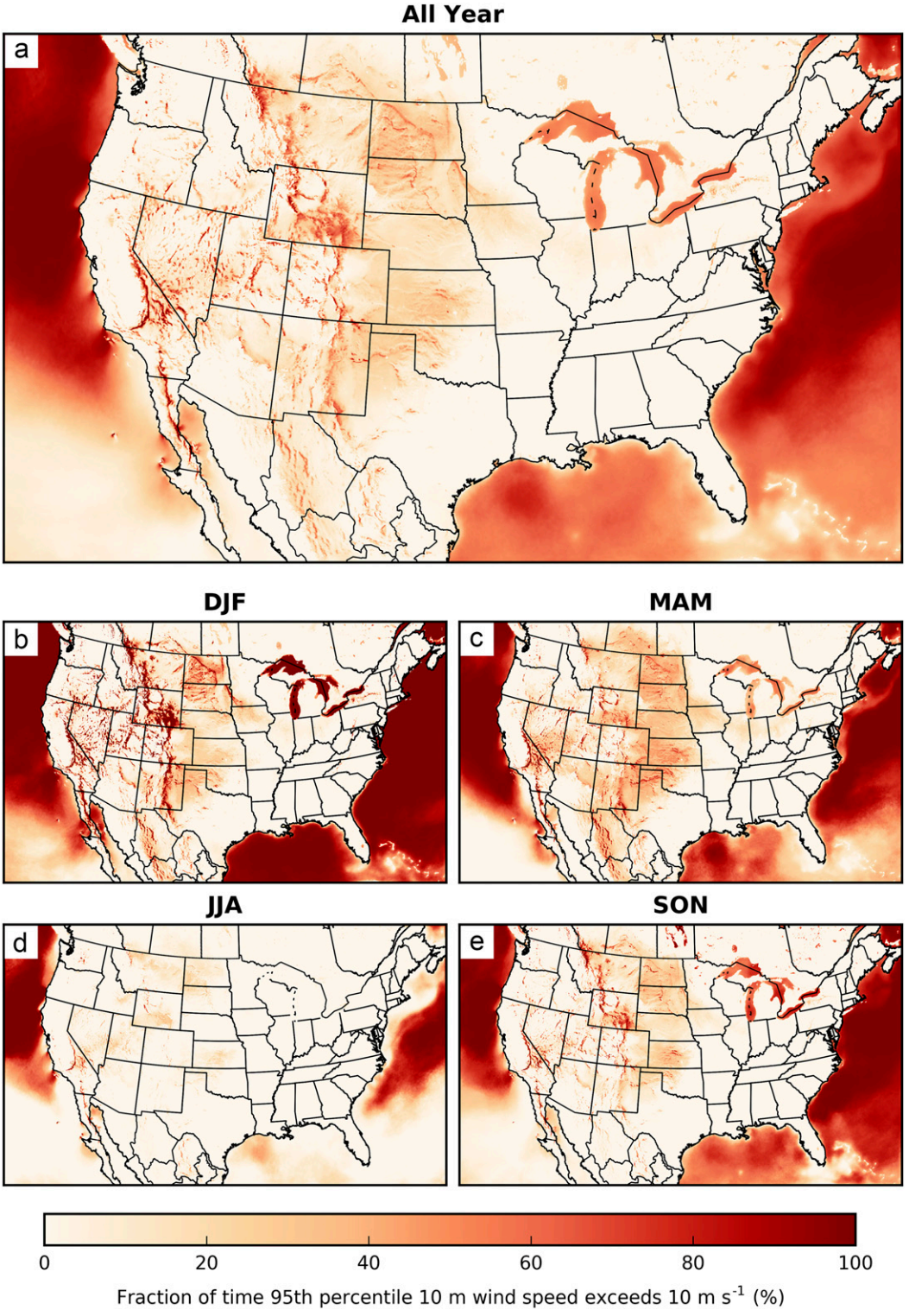


FIG. 1. Fraction of calendar-year hours when the 95th percentile of HRRR 10-m wind speed exceeds 10 m $s^{-1}$ shaded according to the scale during (a) the entire year, (b) winter, (c) spring, (d) summer, and (e) fall.

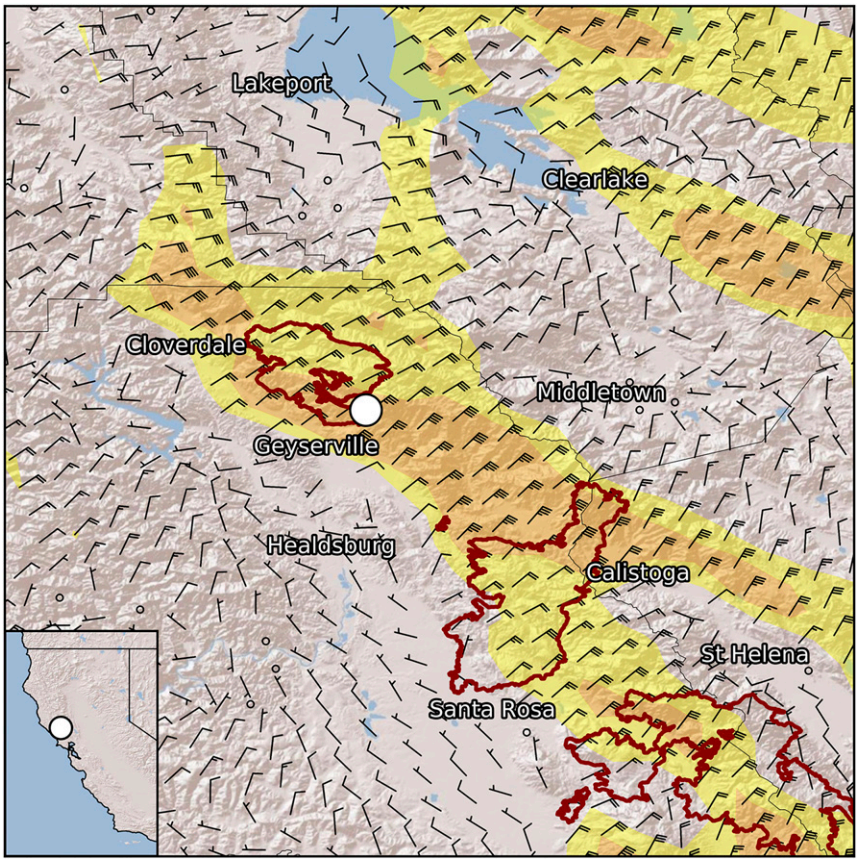


FIG. 2. F00 HRRR 10-m vector winds for 0600 UTC 9 Oct 2017 at every model grid point with half and full barbs denoting 2.5 and 5 m $s^{-1}$, respectively. Yellow (orange) shading indicates areas where wind speeds are greater than 10 (15) m $s^{-1}$. Location of HWKC1 (white dot) and the final fire perimeters (red outlines) for the Pocket (north of Geyserville), Tubbs (north of Santa Rosa), and Nuns (east of Santa Rosa) fires.

calendar-year hour are computed from 90-member samples from April through November derived from the grids for that specific hour from 15 days prior to 14 days after that day during each of the 3 years. (Only 60-member samples between December and March are available at this time because of a disk hardware failure in November 2017.) Hence, we are low-pass smoothing the statistics over 30-day periods but retaining differences from hour to hour to help diagnose diurnal variations that usually dominate weather conditions near the surface. We computed for each sample the mean and 19 different percentiles: 0th (minimum), 1st, 2nd, 3rd, 4th, 5th, 10th, 25th, 33rd, 50th (median), 66th, 75th, 90th, 95th, 96th, 97th, 98th, 99th, and 100th (maximum). Given our sample size of ~100, the finer resolution near the tails of the distribution helps identify the lowest and highest approximately six extreme values.

With the high-throughput design of the OSG, we have efficiently calculated the cumulative distributions for the seven variables at all grid points and each hour of the year and then repeated those calculations multiple times as needed. The work presented here used 8784 unique jobs for each variable—one job for every hour of the year. Each job runs the same script on different OSG remote workers, but each worker node handled the input for a different hour of the year. All jobs were run within a Singularity container set up with 6 GB of memory and the necessary Python software to download and read the HRRR files stored in Gridded Binary, version 2 (GRIB2), format from the Pando archive. The download process was slightly sped up by utilizing Python's multiprocessing module to spin up multiple download instances on all the available cores on the worker node. After a sample of model grids was successfully downloaded, a worker node calculated the mean and the 19 percentiles for every grid point in the model domain. Upon job completion, the statistical output was transferred from the OSG worker node to the OSG home node and then transferred to our local compute resources via Globus bulk file transfer for

analysis. This extra data transfer could have been avoided if our local computer facility was set up as an OSG endpoint, which is planned in the future.

With these cumulative distributions at each model grid point, we can provide information that may be useful for a forecaster faced with evaluating how unusual forecasted conditions may be relative to prior analyses from that model. Instead of relating forecasted values to record minima or maxima, or departures from the analyzed mean, a simple alternative is to compare the forecasted values to the 5th or 95th percentiles. This method helps to compensate for the relatively small sample sizes available to us at this time; helps to mitigate for the highly skewed nature of some fields, such as wind speed; and helps to lessen local differences arising from terrain and land-use variations.

### c. *Statistical calculations from observational data*

We compare the 3-yr HRRR analysis cumulative distributions to observed cumulative distributions obtained from Synoptic Data for observation sites in the vicinity of the North Bay fires. Building on the data collection, archiving, and dissemination tools developed over the years by MesoWest (Horel et al. 2002), Synoptic Data provides application programming interface (API) services to access current and retrospective data that have passed data quality checks from over 45 000 North American locations during the period from 2000 to the present. The approach described earlier to compute cumulative distributions for the HRRR fields was based on the approach developed for Synoptic Data to identify and remove nonphysical observational outliers as a function of locale, time of year, and time of day.

For stations that remained in the same location from as early as 2000 through 2017, observations that passed range, rate-of-change, persistence, and proximity quality-control data checks for a specific hour of a calendar day are collected into 31-day samples centered on that calendar day for the following parameters: temperature, dewpoint temperature, relative humidity, wind speed, wind gust, pressure, altimeter setting, and sea level pressure. The 19 interim percentiles are first derived from samples ranging in size from 93 values (if only 3 years of data are available and if the station reports once per hour) to over 6000 values (if data are available at 5-min intervals during the entire 18-yr period). Since some nonphysical outliers are often still present after those data checks are applied, the differences between the 1st–5th and 95th–99th percentiles are computed, multiplied by a parameter-dependent factor between 2 and 3, and subtracted from (added to) the 5th (95th) percentiles. Then, data falling beyond those new thresholds are removed and the final set of percentiles are created from the trimmed samples. Over 40 billion percentile data values are accessible via the API percentile service.

FIG. 3. HWKC1 RAWS station with Matt Mehle, a NWS incident meteorologist, in the foreground. Photo used with permission.

The following API request returns the minimum ($0\,m\,s^{-1}$), median ($2.2\,m\,s^{-1}$), 95th-percentile ($6.7\,m\,s^{-1}$), and maximum ($18.8\,m\,s^{-1}$) values of wind speed at HWKC1 for 0600 UTC 9 October based on a 551-member sample of observations at that hour from 24 September to 23 October during the 18 years from 2000 to 2017: http://api.synopticdata.com/v2/percentiles?&token=demotoken&start=100906&end=100906&vars=wind_speed&stid=HWKC1&percentiles=0,50,95,100.

## 4. Results and discussion

For model validation and other applications, summary statistics of HRRR model analysis fields at 3-km grid spacing computed over the year or by season are useful metrics. Similar to an evaluation of potential renewable energy resources nationwide by James et al. (2017, 2018), Fig. 1a highlights where the 95th percentiles of the F00 HRRR 10-m wind speed exceed $10\,m\,s^{-1}$ frequently during the year. As should be expected, the areas likely to experience strong wind speeds are over

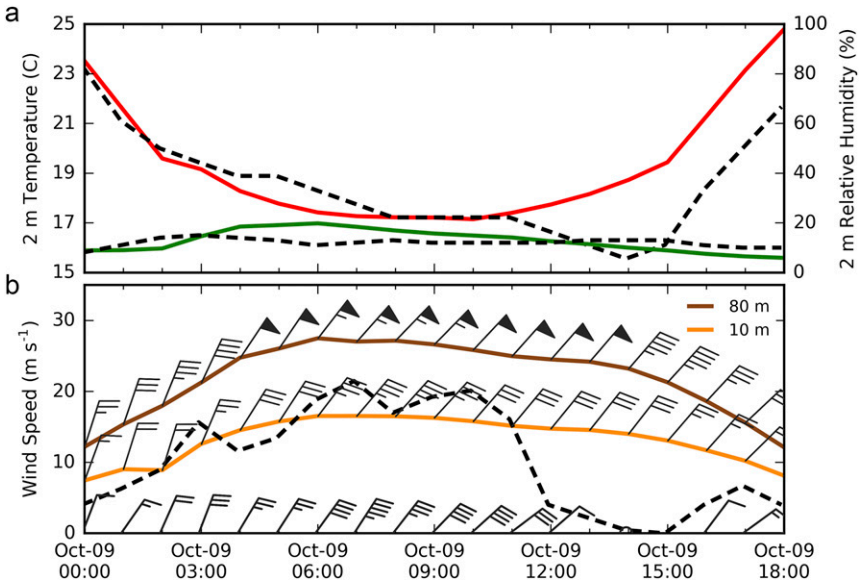


FIG. 4. (a) F00-F18 HRRR forecasts initialized at 0000 UTC 9 Oct 2017 for temperature (red) and relative humidity (green) for the model grid point nearest HWKC1. Temperature and relative humidity observed at HWKC1 during this period denoted by the upper and lower dashed curves, respectively. (b) As in (a), but for HRRR 80-m wind speed (brown), HRRR 10-m wind speed (orange), and HWKC1 wind speed (dashed), including wind vectors with speeds of 2.5, 5, and 25 m $s^{-1}$ delineated by half barbs, full barbs, and flags, respectively.

the oceans, western mountains, the Great Lakes, and the Great Plains (from North Dakota through the Texas Panhandle). Seasonal differences are evident in the other panels of Fig. 1 with high wind speeds over the Great Lakes and western mountains during winter (Fig. 1b) and fall (Fig. 1e). During spring (Fig. 1c), the highest wind speeds over land are analyzed to occur more frequently across the Great Plains. Winds exceeding 10 m $s^{-1}$ are less frequent during summer across the domain (Fig. 1d).

To help motivate the large volume of statistics derived as part of this study, the remainder of this section focuses on illustrating how the cumulative distributions computed across the contiguous United States can be applied regionally and locally to specific wildland fire cases or other extreme weather situations.

### *a. October 2017 North Bay fires*

We focus initially on the 8–9 October 2017 period, when the North Bay fires were ignited north of the San Francisco Bay area. These fires were the costliest and deadliest in California history. They destroyed nearly 7000 structures and caused over 30 deaths. Three of the fires are also referred to as the Central Lake Napa Unit (LNU) complex, consisting of the Pocket (70 $km^2$), Tubbs (149 $km^2$), and Nuns (229 $km^2$) fires. Red flag warnings for the area were issued by the NWS preceding this outbreak. Figure 2 shows the F00 HRRR 10-m vector winds at 0600 UTC 9 October 2017 in the vicinity of the fires. It was near this time that an unusually strong mountain-wave-induced northeasterly winds over and in the lee of the mountain ranges knocked trees into power lines that ignited the fires and wind spread them rapidly (California Department of Forestry and Fire Protection 2018).

We compare the F00 HRRR 10-m winds to those observed at 6.6-m height at the RAWS HWKC1 site, which is located along a grassy ridge near where the Pocket fire began (Fig. 3). Horel and Dong (2010) summarize the reporting characteristics of RAWS intended for fire weather applications relative to those of the NWS Automated Surface Observing Network. Even with the lower sensor height and 10-min reporting interval of HWKC1, this station is quite representative of the wind conditions during this event.

The F00–F18 HRRR forecasts initialized at 0000 UTC 9 October 2017 predicted wind speeds exceeding 10 m $s^{-1}$ for most of the next 18 h, temperatures dropping during the middle of the night, and relative humidity below 20% at the HRRR model grid point nearest HWKC1 (Fig. 4). The highest wind speed (21.5 m $s^{-1}$) at HWKC1 occurred at 0659 UTC at which time the HRRR predicted a 10-m wind speed of 16.5 m $s^{-1}$. After 1100 UTC, the observed wind speed fell below 5 m $s^{-1}$ and temperatures dropped until after sunrise (1400 UTC), but the HRRR 10-m winds remained above 10 m $s^{-1}$ and the 80-m winds exceeded 20 m $s^{-1}$ until after 1600 UTC.

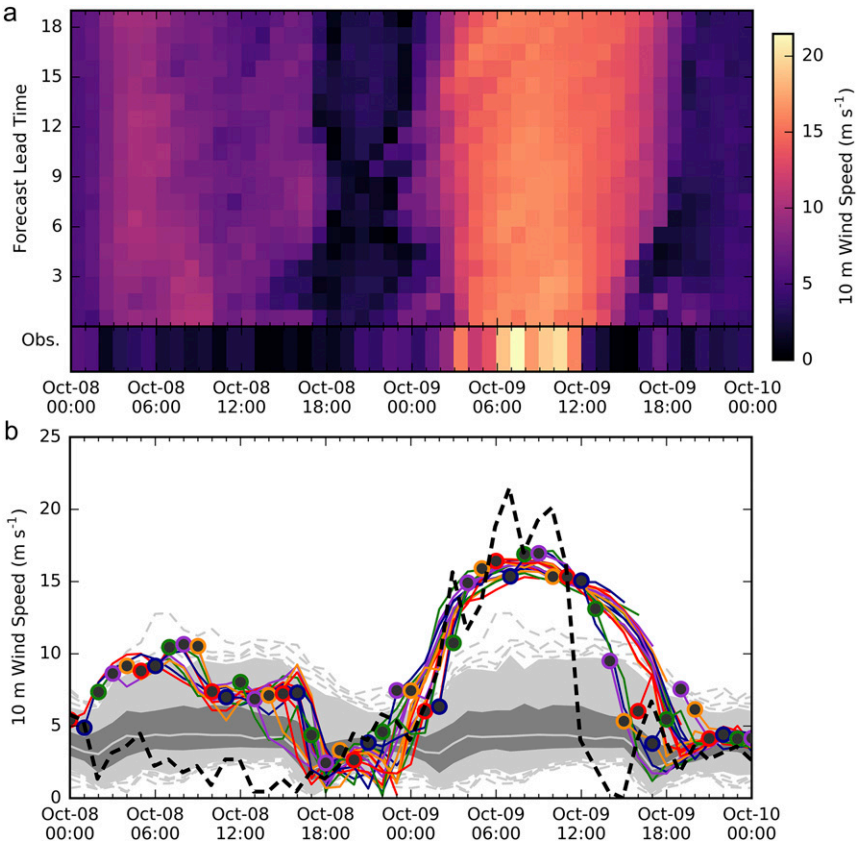


FIG. 5. (a) F00-F18 HRRR 10-m wind speeds shaded according to the accompanying scale at HWKC1 at each forecast lead time as a function of valid time from 0000 UTC 8 Oct 2017 to 0000 UTC 10 Oct 2017. HWKC1 observed wind speeds during this period are shaded at the bottom. (b) Observed HWKC1 wind speeds (heavy dashed line), F00 HRRR wind speeds every hour (gray circles), and F01-F18 HRRR wind speed forecasts initialized every hour (colored lines). Percentiles of HRRR wind speed at HWKC1: median (inner gray line); 25th–75th range (dark gray shading); 5th–95th range (light gray shading); and 1st, 2nd, 3rd, 4th, 96th, 97th, 98th, and 99th percentiles (dashed gray lines).

The Hovmöller-style diagram in Fig. 5a and the line graphics in Fig. 5b highlight the run-to-run consistency of the HRRR winds. Each model run predicted the onset of the strong winds between 0400 and 1200 UTC but sustained the strong winds in the vicinity of HWKC1 later into the following morning than observed. The shading and dashed gray lines in Fig. 5b indicate the F00 HRRR wind percentiles computed separately for each hour using the OSG. The forecasted and observed wind speeds at 0600 UTC 9 October were unusual for this location and time of year, as these winds exceeded the 3-yr 99th-percentile wind speed. (This exceptional event is included in the 2015–17 sample from which the percentiles were computed.) It is also clear that these excessive winds are unusual during the month of October from the 18-yr sample of observed wind speeds at HWKC1, where winds exceeding $10\,m\,s^{-1}$ account for only ~1% of the time (Fig. 6a). In fact, the maximum observed wind during the event, $21\,m\,s^{-1}$, was the highest recorded wind speed for that hour and day at HWKC1 for the period of record.

Comparisons of the observed versus modeled cumulative distributions at HWKC1 are one means to highlight model biases (Wilks 2011; Mittermaier and Csima 2017). We focus on 0600 UTC, which is near the time of maximum winds for this event. In general, F00 HRRR winds in October at 0600 UTC for the grid point nearest HWKC1 within the more limited 3-yr sample tend to exceed those observed (Fig. 6a), although they are more comparable in Fig. 6b when extending the comparison to a larger sample of the nearby model grid points within a 33 km × 33 km box centered on HWKC1. There are clear diurnal differences at this location evident in Fig. 6c with the tendency for stronger HRRR winds at

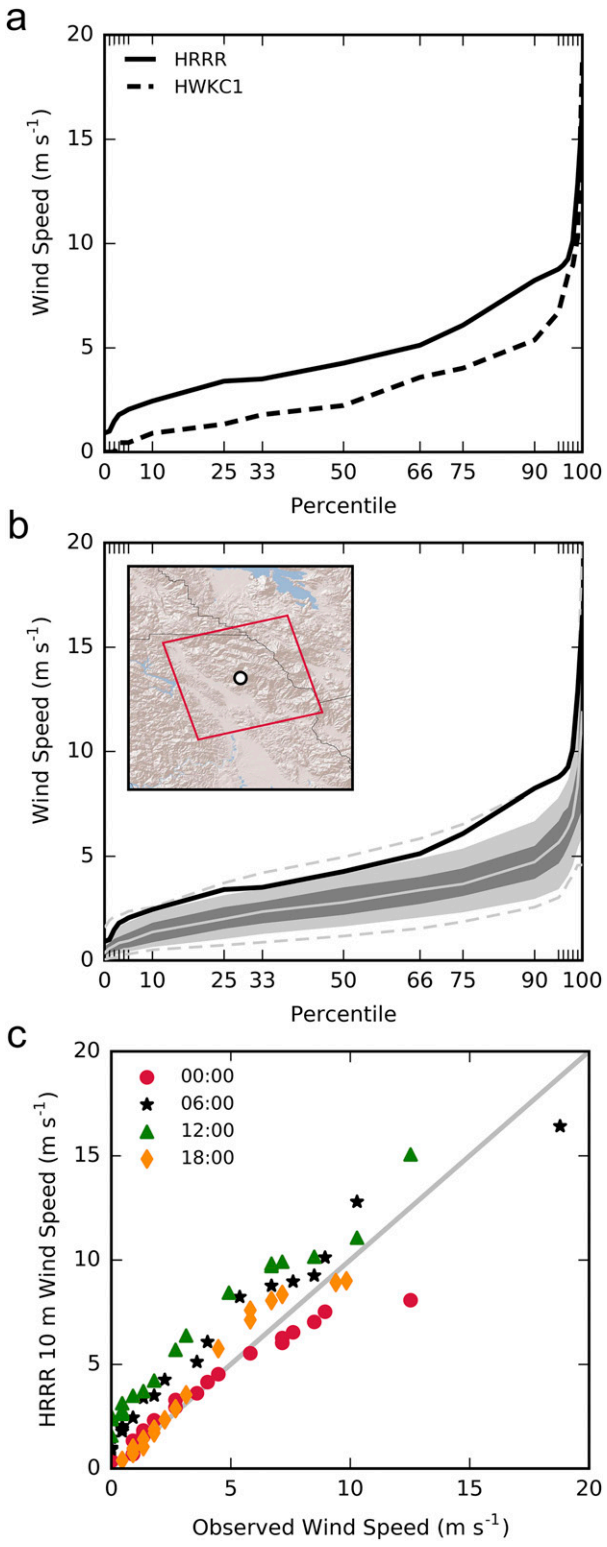


FIG. 6. (a) Wind speed cumulative distributions at HWKC1 at 0600 UTC 9 Oct from the 18-yr observed sample (dashed) and 3-yr sample F00 HRRR sample (solid). (b) As in (a), but summarized for the sample of 121 nearby grid points within the red box by the median (inner gray line), 25th–75th quartiles (dark gray shading), 5th–95th percentiles (light gray shading), and minima and maxima (light gray dashed lines). The solid black line is repeated from (a). (c) Comparison of the 19 percentile values (from minima to maxima) at HWKC1 for the observed and F00 HRRR wind speeds at 0000 (red circles), 0600 (black stars), 1200 (green triangles), and 1800 UTC (orange diamonds) 9 Oct.

0600 and 1200 UTC and more comparable distributions during late morning and afternoon (1800 and 0000 UTC, respectively).

The cumulative distribution for any HRRR output variable at any point in the domain varies by season and time of day. Figure 7 summarizes distributions of F00 HRRR 2-m air temperature and 10-m wind speed for every day of the year restricted to 0600 UTC for the grid point in the HRRR domain nearest HWKC1 relative to those observed at HWKC1. As might be expected based on the relative sample sizes, the spread between the 5th and 95th percentiles of observed air temperature is wider than that from the much smaller 3-yr HRRR sample during nearly every day of the year. The tendency for nearly the same observed median wind speeds at this time of the night results from the RAWS standard to report wind speeds as integer miles-per-hour values. Overall, there is a tendency for the HRRR F00 wind speeds to be higher than those observed at HWKC1. Additional information on the typical conditions for an area that can be summarized in terms of the median or the inner quartile range may also have value to a forecaster.

The comparisons between observed and HRRR cumulative distributions in Figs. 6 and 7 are useful to identify general tendencies and limitations for both the observations and the model fields. As mentioned in section 3b, of greater interest to us is to develop a foundation on which to enhance situational awareness of developing extreme events. For example, Fig. 8 illustrates differences in the F18 and F12 10-m vector wind forecasts valid 0600 UTC 9 October 2017 and the observed and HRRR 95th percentiles for the same domain used in Fig. 2. The shaded circles in Figs. 8a and 8c illustrate that the winds at many locations, particularly those along ridges, were forecasted to be 8–14 m s$^{-1}$ faster than the observed 95th-percentile values at those locations during the middle of the night on 9 October. Relative to the F00 HRRR wind 3-yr climatology (Figs. 8b and 8d), anomalous bands of forecasted high winds over and in the lee of the local high terrain are evident.

### b. *April 2018 western Oklahoma fires*

An April 2018 fire outbreak in western Oklahoma, which is independent of the 2015–17 climatological

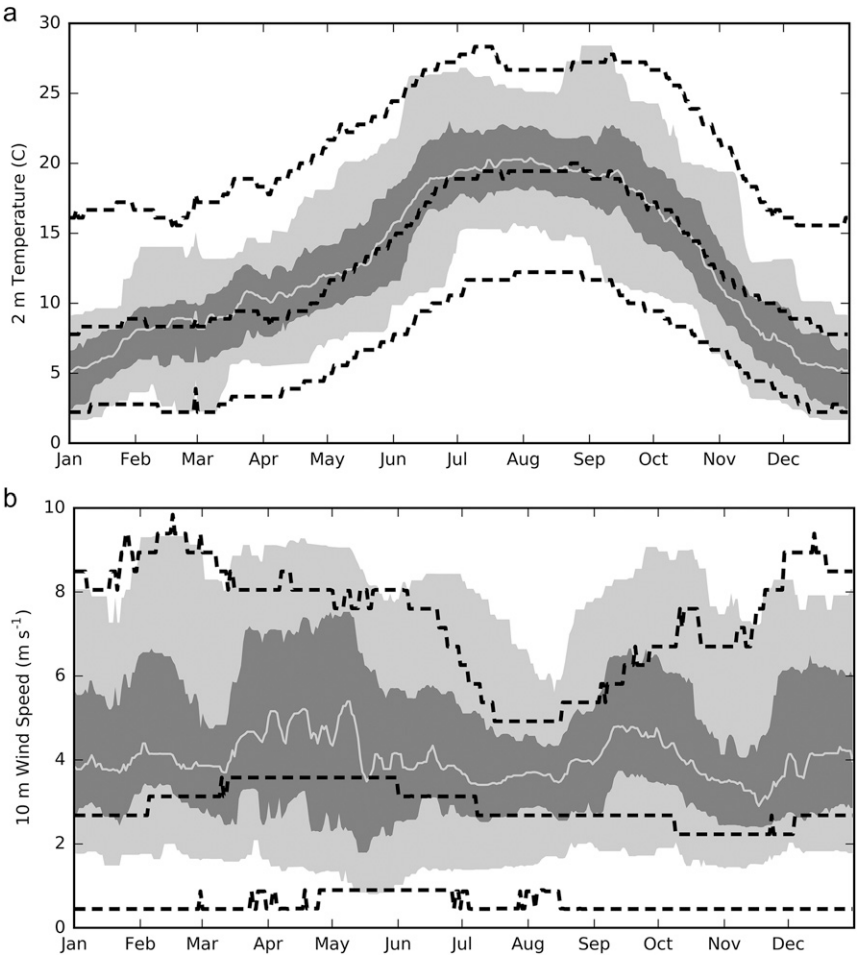


FIG. 7. Percentiles for 0600 UTC for each day of the climatological year at HWKC1 for (a) 2-m temperature and (b) wind speed. Observed 5th, 50th (median), and 95th percentiles denoted by dashed lines and F00 HRRR percentiles depicted as follows: median, center gray line; 25th–75th quartiles, dark gray shading; and 5th–95th percentiles, light gray shading.

period, is used here as an additional case study. The NWS issued a red flag warning and the Oklahoma Forestry Service posted extreme fire danger for western Oklahoma for 12 April 2018, as a synoptic-scale trough was expected to deepen and lead to hot, dry, windy conditions across the region. The Rhea fire in western Oklahoma began at 1900 UTC 12 April 2018 and grew explosively during the next day before being fully contained at over 1100 $km^2$ later in the month. The extreme fire temperatures and centermost large smoke plume in Fig. 9 are associated with the Rhea fire. Figure 10 highlights the rapid deepening and progression of the upper-level trough evident in the 18-h HRRR forecasts of 500-hPa geopotential height valid at 0000 and 1500 UTC 13 April 2018.

Figure 10a highlights the areas where the 18-h forecast of 2-m temperature valid at 0000 UTC 13 April 2018 was greater than the HRRR 95th percentiles and less than the 5th percentiles of 2-m temperature for that hour and day of the calendar year. Ahead of the upper-level trough, air temperatures in Kansas, western Oklahoma, and western Texas were forecasted to be over 6°C higher than the 95th-percentile temperatures, while the forecasted temperatures were expected to be below the 5th-percentile temperatures by 2°–4°C from Utah extending northward into Canada. The areas of large air temperature departures for the forecast valid at 1500 UTC 13 April shifted eastward with the progression of the upper-level trough (Fig. 10b) and were accompanied with negative departures of 2-m dewpoint temperature approaching 10°C in western Oklahoma and Texas (Fig. 10d). Forecasted wind speeds above the 95th percentile were widespread ahead of and behind the trough (Figs. 10e and 10f). Comparing the

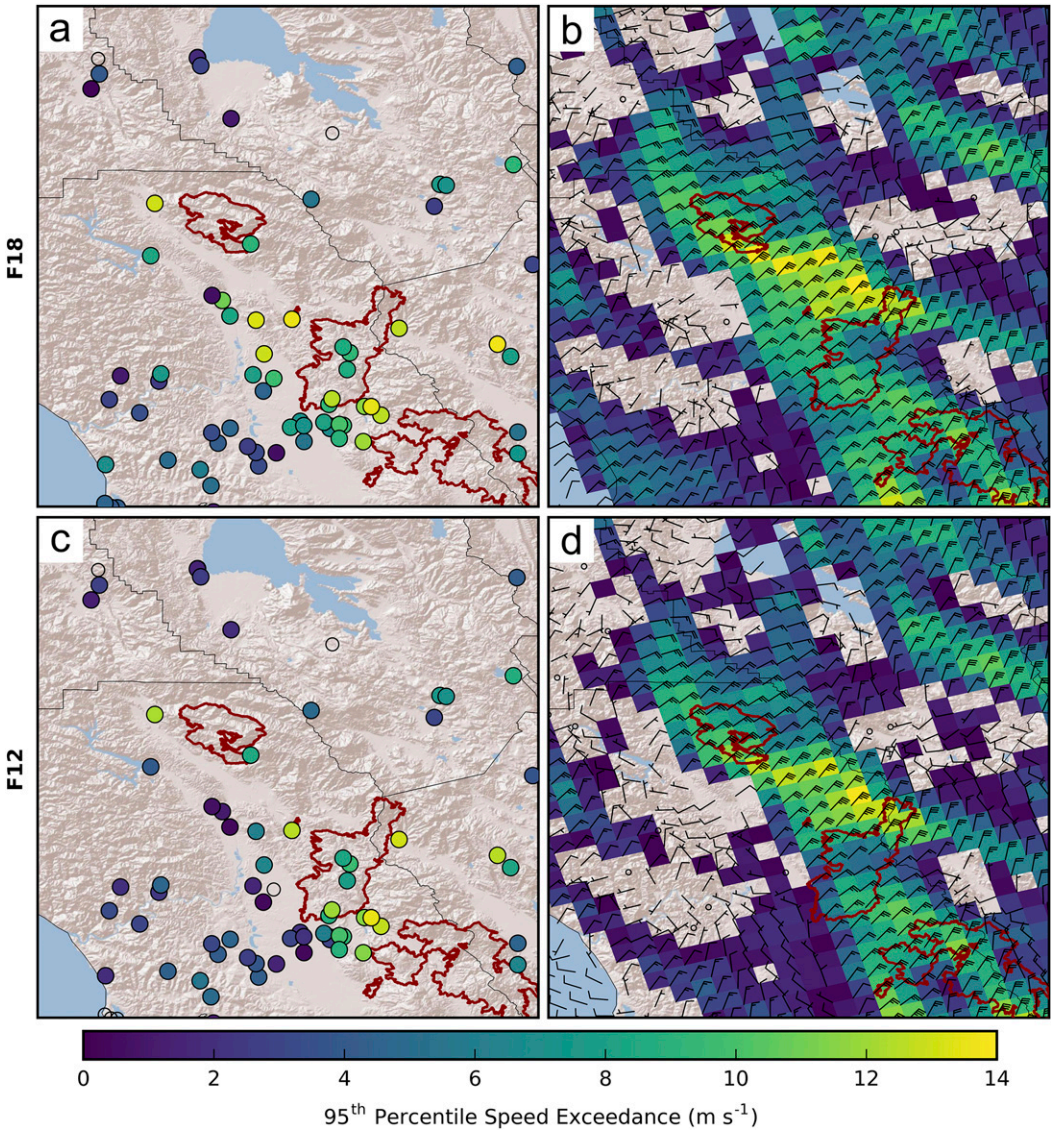


FIG. 8. (a) Differences between F18 HRRR 10-m wind speed valid 0600 UTC 9 Oct 2017 and the 95th percentile of observed wind speeds at 0600 UTC 9 Oct shaded according to the scale at the bottom. Empty circles denote stations where the forecasted wind did not exceed the 95th percentile for that hour and day of the calendar year. (b) As in (a), but for the forecast relative to the 95th-percentile F00 HRRR wind speeds. Unshaded regions reflect locations where the forecasted wind did not exceed the 95th percentile. F18 HRRR 10-m vector winds plotted as in Fig. 2. (c) As in (a), but for F12 HRRR 10-m wind speed valid 0600 UTC 9 Oct 2017. (d) As in (b), but for F12 HRRR 10-m wind speed valid 0600 UTC 9 Oct 2017.

forecasts to the cumulative distributions on the national level in this manner gives forecasters additional insight into how the forecasts relate to conditions observed during recent years for the region.

## 5. Summary

We have demonstrated how the OSG has been used to efficiently compute a set of cumulative distributions for selected variables from HRRR F00 analyses (F00) over a 3-yr period. The HRRR model output at hourly intervals and 3-km grid spacing provides an opportunity to examine near-surface weather conditions throughout the contiguous United States that is not possible from current observational resources. The cumulative distributions for this high-temporal- and high-spatial-resolution model may then be used to estimate typical and atypical atmospheric conditions at any location within the HRRR

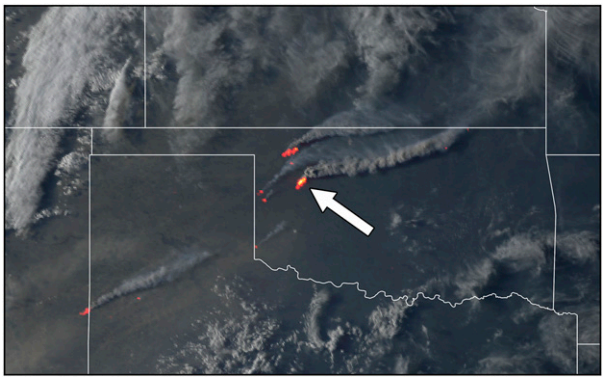

FIG. 9. *GOES-16* true color and fire temperature blend at 2342 UTC 12 Apr 2018. Arrow points to the Rhea fire.

domain as a function of hour of the calendar year. Although the examples presented here are affected to some extent by the limited 3-yr sample size available at this time, we intend to recompute these cumulative distributions annually as the available period of record lengthens. Distributing the total work across the OSG made it possible to compute and recompute multiple times hourly statistics for several variables using far less computer time than would have been required on our local compute nodes and freed those nodes up for other work.

In hindsight, we realize the total number of jobs and time to complete the work could have been significantly reduced had each worker node been assigned to perform the statistical calculation for 2 or more hours and utilized the grids it had already downloaded. Future use of the OSG will follow this approach to reduce the total input/output (I/O) required to complete the work, thus allowing for more variables to be analyzed in the same amount of time. Other users of the OSG should also consider I/O, memory, and disk requirements to optimize their jobs.

The computational resources provided by the OSG are an underutilized resource in environmental fields for analyzing massive datasets. The need for data analytics applied efficiently to rapidly expanding resources in the atmospheric sciences is growing, for example, *GOES-16* and *GOES-17* imagery and model output from NOAA's unified Environmental Modeling System, which includes the Next Generation Global Prediction System available in the next several years. We illustrated how a suite of statistical calculations can be derived efficiently using the OSG on samples of analyses from the high-spatial- and high-temporal-resolution HRRR model across the multiple dimensions of variable, location, and time within the calendar year.

Our objective was to develop a foundation on which to use historical model data to highlight typical and atypical weather conditions that impact wildland fire conditions throughout the contiguous United States. Two extreme wildfire cases, North Bay fires in California and the Rhea fire in Oklahoma, were used to demonstrate the potential utility of empirical cumulative distributions of hourly HRRR analyses. Awareness of how unusual present and forecasted weather conditions are near wildland fires may assist fire suppression efforts and firefighter safety. Wildland fires generally occur in remote areas, where incident meteorologists assigned to large fires and office-based NWS forecasters issuing spot forecasts are likely unfamiliar with forecasting for the specific area of a fire or evaluating how a particular model behaves for that area and time of year. The capabilities developed here are intended to extend web-based capabilities developed to monitor weather conditions in the Great Lakes and Alaska regions for wildland fire applications (Horel et al. 2014; https://glff.mesowest.org/; https://akff.mesowest.org).

As with any statistical analysis of operational models that undergo periodic upgrades, the specific results presented here should be treated with some caution, as differing biases in model performance may be evident before and after model upgrades. Particular concern for this study was the release in August 2016 of version 2 of the HRRR model for the contiguous United States. However, the comparisons of the observed cumulative distributions at HWKC1 and nearby HRRR analysis grid points provide confidence that the HRRR samples are not completely overwhelmed by either the small sample size or changes resulting from the upgrade.

Palutikof et al. (1999) describe several methods to estimate extreme wind quantiles from samples limited to 2–3 years. We illustrated one of their recommended approaches qualitatively (Fig. 6b), that is, to artificially increase the sample size used to compute cumulative distributions by comparing neighboring locations. We obviously are more successful capturing the seasonal variations in typical values as a function of time of day than extreme values in locales dominated by large interannual variations. Furthermore, comparing the departures of forecast fields relative to the cumulative distributions calculated for each field, as done in Figs. 8 and 10 for the 5th and 95th percentiles, reduces local variations resulting from differences in terrain and land surface characteristics in a similar way to comparing forecasts against standardized anomalies. As might be expected, our two extreme examples of the North Bay and Rhea fires yield large departures from those thresholds locally as well as regionally.

Future work will benefit from an extended record of HRRR data, including the version 3 upgrade that occurred on 12 July 2018. At that time, the Alaska HRRR

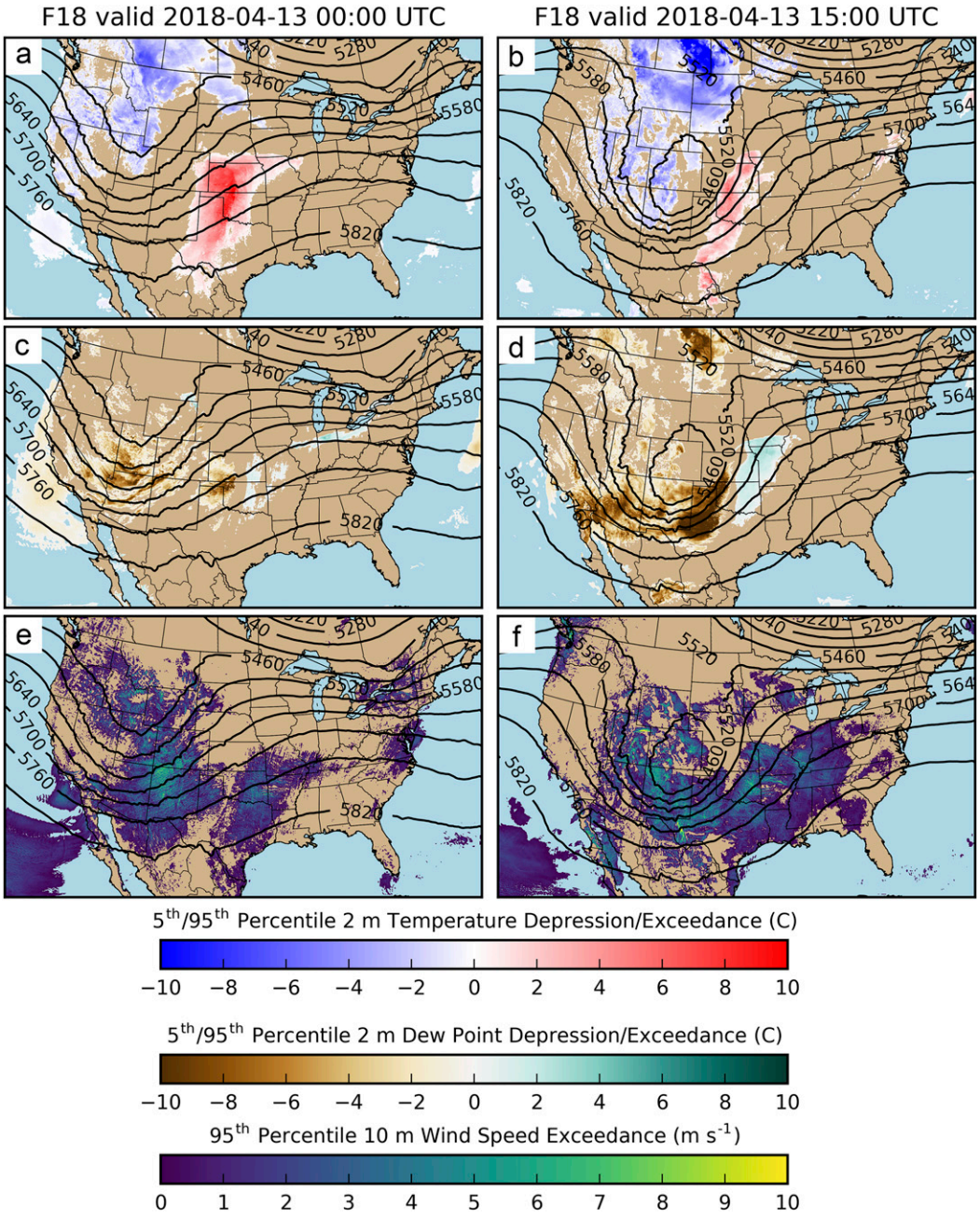


FIG. 10. (a) F18 500-hPa geopotential height valid 0000 UTC 13 Apr 2018 (black contours at 40-m intervals). Negative (positive) departures between F18 HRRR 2-m temperature valid 0000 UTC 13 Apr 2018 and 5th (95th)-percentile F00 HRRR temperature at 0000 UTC 13 Apr shaded according to the appropriate scale at the bottom. (b) As in (a), but valid 1500 UTC 13 Apr 2018. (c) As in (a), but for 2-m dewpoint temperature. (d) As in (b) but for 2-m dewpoint temperature. (e) As in (a), but for 10-m wind speed positive departures only. (f) As in (b), but for 10-m wind speed positive departures only.

model also became operational (McCorkle et al. 2018; https://rapidrefresh.noaa.gov/hrrr/). These memory and I/O intensive computations could be extended to over 100 additional meteorological variables in the HRRR surface files alone and the hundreds of fields available in 25-hPa increments in the vertical and the model forecasts. A further extension to this work could be to compute cumulative distributions for multivariate conditions rather than the univariate results presented here. For example, cumulative distributions of wind speed,

temperature, and relative humidity that exceed or lie below specific thresholds could be used to identify the likelihood of red flag warning conditions throughout the contiguous United States and Alaska as a function of hour in the calendar year, albeit having to factor in the varying regional differences in warning criteria. Also, model forecasts of a new index intended for wildland fire applications (the hot–dry–windy index; McDonald et al. 2018; Srock et al. 2018) could be evaluated in this manner.

An immediate application of the HRRR cumulative distributions will be to flag potentially erroneous station observations archived since 2000 by MesoWest and Synoptic Data. The iterative approach outlined in section 3c is intended to be extended to exclude observed values that deviate far from the tails of the HRRR analysis values near station locations. This work and other data analysis tasks are scalable and can be accomplished with the high-throughput computing resource from OSG.

*Acknowledgments.* This work was funded by the Joint Fire Science Program Grant L17AC00225, National Science Foundation Grant 1443046, and Synoptic Data Contract PO17-00640. We appreciate the support and resources from the University of Utah Center for High Performance Computing. We thank the Climate Corporation for backfilling portions of our NOAA HRRR archive prior to 2018. In addition, this research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation Award 1148698, and the U.S. Department of Energy's Office of Science.

## REFERENCES

Ansari, S., and Coauthors, 2018: Unlocking the potential of NEXRAD data through NOAA's Big Data Partnership. *Bull. Amer. Meteor. Soc.*, **99**, 189–204, https://doi.org/10.1175/BAMS-D-16-0021.1.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Blaylock, B., J. Horel, and S. Liston, 2017a: Cloud archiving and data mining of High-Resolution Rapid Refresh forecast model output. *Comput. Geosci.*, **109**, 43–50, https://doi.org/10.1016/j.cageo.2017.08.005.

——, ——, and E. Crosman, 2017b: Impact of lake breezes on summer ozone concentrations in the Salt Lake Valley. *J. Appl. Meteor. Climatol.*, **56**, 353–370, https://doi.org/10.1175/JAMC-D-16-0216.1.

California Department of Forestry and Fire Protection, 2018: CAL FIRE investigators determine causes of 12 wildfires in Mendocino, Humboldt, Butte, Sonoma, Lake, and Napa Counties. CAL FIRE News Release, 2 pp., http://www.calfire.ca.gov/communications/downloads/newsreleases/2018/2017_WildfireSiege_Cause.pdf.

Chapman, L., and S. J. Bell, 2018: High-resolution monitoring of weather impacts on infrastructure networks using the Internet of Things. *Bull. Amer. Meteor. Soc.*, **99**, 1147–1154, https://doi.org/10.1175/BAMS-D-17-0214.1.

Chekanov, S. V., I. Pogrebnyak, and D. Wilbern, 2017: Cross-platform validation and analysis environment for particle physics. *Comput. Phys. Commun.*, **220**, 91–96, https://doi.org/10.1016/j.cpc.2017.06.017.

Foster, I., 2005: Service-oriented science. *Science*, **308**, 814–817, https://doi.org/10.1126/science.1110411.

Gravelle, C. M., K. J. Runk, K. L. Crandall, and D. W. Snyder, 2016: Forecaster evaluations of high temporal satellite imagery for the GOES-R era at the NWS Operations Proving Ground. *Wea. Forecasting*, **31**, 1157–1177, https://doi.org/10.1175/WAF-D-15-0133.1.

Hoffman, R. N., S. Boukabara, V. K. Kumar, K. Garrett, S. P. Casey, and R. Atlas, 2017: An empirical cumulative density function approach to defining summary NWP forecast assessment metrics. *Mon. Wea. Rev.*, **145**, 1427–1435, https://doi.org/10.1175/MWR-D-16-0271.1.

Horel, J. D., and X. Dong, 2010: An evaluation of the distribution of Remote Automated Weather Stations (RAWS). *J. Appl. Meteor. Climatol.*, **49**, 1563–1578, https://doi.org/10.1175/2010JAMC2397.1.

——, and Coauthors, 2002: MesoWest: Cooperative mesonets in the western United States. *Bull. Amer. Meteor. Soc.*, **83**, 211–226, https://doi.org/10.1175/1520-0477(2002)083<0211:MCMITW>2.3.CO;2.

——, R. Ziel, C. Galli, J. Pechmann, and X. Dong, 2014: An evaluation of fire danger and behavior indices in the Great Lakes region calculated from station and gridded weather information. *Int. J. Wildland Fire*, **23**, 202–214, https://doi.org/10.1071/WF12186.

Houston, A. L., N. A. Lock, J. Lahowetz, B. L. Barjenbruch, G. Limpert, and C. Oppermann, 2015: Thunderstorm Observation by Radar (ThOR): An algorithm to develop a climatology of thunderstorms. *J. Atmos. Oceanic Technol.*, **32**, 961–981, https://doi.org/10.1175/JTECH-D-14-00118.1.

Huerta, E. A., and Coauthors, 2017: BOSS-LDG: A novel computational framework that brings together blue waters, open science grid, shifter and the LIGO Data Grid to accelerate gravitational wave discovery. *2017 IEEE 13th International Conference on e-Science (e-Science)*, IEEE, 335–344, https://doi.org/10.1109/eScience.2017.47.

Hwang, Y., A. J. Clark, V. Lakshmanan, and S. E. Koch, 2015: Improved nowcasts by blending extrapolation and model forecasts. *Wea. Forecasting*, **30**, 1201–1217, https://doi.org/10.1175/WAF-D-15-0057.1.

James, E. P., and S. G. Benjamin, 2017: Observation system experiments with the hourly updating Rapid Refresh model using GSI hybrid ensemble–variational data assimilation. *Mon. Wea. Rev.*, **145**, 2897–2918, https://doi.org/10.1175/MWR-D-16-0398.1.

——, ——, and M. Marquis, 2017: A unified high-resolution wind and solar dataset from a rapidly updating numerical weather prediction model. *Renewable Energy*, **102**, 390–405, https://doi.org/10.1016/j.renene.2016.10.059.

——, ——, and ——, 2018: Offshore wind speed estimates from a high-resolution rapidly updating numerical weather prediction model forecast dataset. *Wind Energy*, **21**, 264–284, https://doi.org/10.1002/we.2161.

Juve, G., M. Rynge, E. Deelman, J.-S. Vockler, and G. Berriman, 2013: Comparing FutureGrid, Amazon EC2, and Open

Science Grid for scientific workflows. *Comput. Sci. Eng.*, **15**, 20–29, https://doi.org/10.1109/MCSE.2013.44.

Loeser, C. F., M. A. Herrera, and I. Szunyogh, 2017: An assessment of the performance of the operational global ensemble forecast systems in predicting the forecast uncertainty. *Wea. Forecasting*, **32**, 149–164, https://doi.org/10.1175/WAF-D-16-0126.1.

McCorkle, T. A., J. D. Horel, A. A. Jacques, and T. Alcott, 2018: Evaluating the experimental High-Resolution Rapid Refresh–Alaska modeling system using USArray pressure observations. *Wea. Forecasting*, **33**, 933–953, https://doi.org/10.1175/WAF-D-17-0155.1.

McDonald, J., A. Srock, and J. Charney, 2018: Development and application of a Hot-Dry-Windy Index (HDW) climatology. *Atmosphere*, **9**, 285, https://doi.org/10.3390/atmos9070285.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Mittermaier, M., and G. Csima, 2017: Ensemble versus deterministic performance at the kilometer scale. *Wea. Forecasting*, **32**, 1697–1709, https://doi.org/10.1175/WAF-D-16-0164.1.

Mülmenstädt, J., D. Lubin, L. M. Russell, and A. M. Vogelmann, 2012: Cloud properties over the North Slope of Alaska: Identifying the prevailing meteorological regimes. *J. Climate*, **25**, 8238–8258, https://doi.org/10.1175/JCLI-D-11-00636.1.

Norman, A., and Coauthors, 2015: Large scale Monte Carlo simulation of neutrino interactions using the Open Science Grid and commercial clouds. *J. Phys.: Conf. Ser.*, **664**, 032023, https://doi.org/10.1088/1742-6596/664/3/032023.

Oswald, E. M., 2018: An analysis of the prevalence of heat waves in the United States between 1948 and 2015. *J. Appl. Meteor. Climatol.*, **57**, 1535–1549, https://doi.org/10.1175/JAMC-D-17-0274.1.

Page, W. G., N. S. Wagenbrenner, B. W. Butler, J. M. Forthofer, and C. Gibson, 2018: An evaluation of NDFD weather forecasts for wildland fire behavior prediction. *Wea. Forecasting*, **33**, 301–315, https://doi.org/10.1175/WAF-D-17-0121.1.

Palutikof, J. P., B. B. Brabson, D. H. Lister, and S. T. Adcock, 1999: A review of methods to calculate extreme wind speeds. *Meteor. Appl.*, **6**, 119–132, https://doi.org/10.1017/S1350482799001103.

Poehlman, W., M. Rynge, C. Branton, D. Balamurugan, and F. Feltus, 2016: OSG-GEM: Gene expression matrix construction using the Open Science Grid. *Bioinform. Biol. Insights*, **10**, 133–141, https://doi.org/10.4137/BBI.S38193.

Pordes, R., and Coauthors, 2007: The Open Science Grid. *J. Phys.: Conf. Ser.*, **78**, 012057, https://doi.org/10.1088/1742-6596/78/1/012057.

——, and Coauthors, 2008: New science on the Open Science Grid. *J. Phys.: Conf. Ser.*, **125**, 012070, https://doi.org/10.1088/1742-6596/125/1/012070.

Powers, J., and Coauthors, 2017: The Weather Research and Forecasting Model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737, https://doi.org/10.1175/BAMS-D-15-00308.1.

Régin, J. C., M. Rezgui, and A. Malapert, 2013: Embarrassingly parallel search. *Principles and Practice of Constraint Programming*, C. Schulte, Ed., CP 2013, Lecture Notes in Computer Science, Vol. 8124, Springer, 596–610, https://doi.org/10.1007/978-3-642-40627-0_45.

Rogers, J. W., R. L. Thompson, and P. T. Marsh, 2014: Potential applications of a CONUS sounding climatology developed at the Storm Prediction Center. *27th Conf. Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 145, https://ams.confex.com/ams/27SLS/webprogram/Paper255385.html.

Schmit, T. J., P. Griffith, M. M. Gunshor, J. M. Daniels, S. J. Goodman, and W. J. Lebair, 2017: A closer look at the ABI on the GOES-R series. *Bull. Amer. Meteor. Soc.*, **98**, 681–698, https://doi.org/10.1175/BAMS-D-15-00230.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

Sfiligoi, I., D. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Wurthwein, 2009: The pilot way to grid resources using glide in WMS. *2009 WRI World Congress on Computer Science and Information Engineering (CSIE 2009)*, Vol. 2, IEEE, 428–432, https://doi.org/10.1109/CSIE.2009.950.

Srock, A., J. Charney, B. Potter, and S. Goodrick, 2018: The Hot-Dry-Windy Index: A new fire weather index. *Atmosphere*, **9**, 279, https://doi.org/10.3390/atmos9070279.

Steward, J. L., A. Aksoy, and Z. S. Haddad, 2017: Parallel direct solution of the ensemble square root Kalman filter equations with observation principal components. *J. Atmos. Oceanic Technol.*, **34**, 1867–1884, https://doi.org/10.1175/JTECH-D-16-0140.1.

Stockinger, H., M. Pagni, L. Cerutti, and L. Falquet, 2006: Grid approach to embarrassingly parallel CPU-intensive bioinformatics problems. *2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, IEEE, 58, https://doi.org/10.1109/E-SCIENCE.2006.261142.

Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.

Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, https://doi.org/10.1175/JCLI-D-16-0652.1.