Money Doesn't Grow on Trees, But Forecasts Do: Forecasting Extreme Precipitation with Random Forests

Gregory R. Herman* and Russ S. Schumacher

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

Corresponding author address: Gregory R. Herman, Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, CO 80523.

E-mail: gherman@atmos.colostate.edu

ABSTRACT

Approximately eleven years of reforecasts from NOAA's Second Generation Global Ensemble Forecast System Reforecast (GEFS/R) model are used to train a contiguous United States (CONUS)-wide gridded probabilistic prediction system for locally extreme precipitation, developed primarily using the random forest (RF) algorithm. Locally extreme precipitation is quantified for 24-hour precipitation accumulations in the framework of average recurrence intervals (ARIs), with two severity levels: 1- and 10-year ARI exceedances. Forecasts are made from 0000 UTC forecast initializations for two 1200 UTC–1200 UTC periods: Days 2 and 3 comprising respectively forecast hours 36–60 and 60–84. Separate models are trained for each of eight forecast regions and for each forecast lead time. GEFS/R predictors vary in space and time relative to the forecast point, and include not only the quantitative precipitation forecast (QPF) output from the model, but also variables that characterize the meteorological regime, including winds, moisture, and instability. Numerous sensitivity experiments are performed to determine the effects of the inclusion or exclusion of different aspects of forecast information in the model predictors, the choice of statistical algorithm, and the effect of performing dimensionality reduction via principal component analysis as a pre-processing step. Overall, it is found that the machine learning (ML)-based forecasts add significant skill over exceedance forecasts produced from both the raw GEFS/R ensemble QPFs and from the European Centre for Medium-Range Weather Forecasts' (ECMWF) global ensemble across almost all regions of CONUS. ML-based forecasts are found to be underconfident, while raw ensemble forecasts are highly overconfident.

1. Introduction

Locally extreme precipitation can cause a variety of costly, disruptive, and endangering impacts, including flooding, flash flooding, and landslides. In 2016 alone, these hazards combined caused more than 120 fatalities and $10 billion in damages over the United States (NWS 2017b). The prediction of flash floods is a notoriously challenging forecast problem, requiring not only accurate prediction of heavy rainfall magnitudes, but also of the spatiotemporal distribution of that rainfall; the hydrologic interactions between precipitation, terrain, and the land surface; and also of antecedent precipitation and its effects on soil conditions. Forecasting precipitation processes responsible for most observed extreme rainfall over the contiguous United States (CONUS) is often considered among the most challenging problems in contemporary numerical weather prediction (NWP; e.g. Fritsch and Carbone 2004; Novak et al. 2014). Given that the rainfall forecast alone presents such a considerable challenge, the additional hydrologic considerations in the flash flood forecast problem present an even more daunting task. While recent advances in heavy rainfall and flash flood forecasting have been made (e.g. Hapuarachchi et al. 2011; Novak et al. 2014; Barthold et al. 2015), forecasts still struggle in many situations (e.g. Delrieu et al. 2005; Lackmann 2013; Schumacher et al. 2013; Gochis et al. 2015; Nielsen and Schumacher 2016, among many others) and substantial progress remains to be made.

Contemporary operational dynamical forecast models often struggle to accurately simulate the physical processes responsible for its production. For example, models with parameterized convection often have a variety of persistent errors and biases associated with their depiction of convective systems, convective systems being responsible for the majority of flooding rains over much of CONUS (e.g. Schumacher and Johnson 2006; Stevenson and Schumacher 2014; Herman and Schumacher 2016a). These include a tendency to underpredict total rainfall from convective sys-

tems (e.g. Schumacher and Johnson 2008; Herman and Schumacher 2016a); produce systems displaced too far to the north and west from where they are observed (e.g. Grams et al. 2006; Wang et al. 2009; Clark et al. 2010); initiate convection too early (e.g. Davis et al. 2003; Wilson and Roberts 2006; Clark et al. 2007); generate systems with too large an areal extent (e.g. Wilson and Roberts 2006); and propagate them incorrectly, too slowly, or not at all (e.g. Davis et al. 2003; Pinto et al. 2015). While convection-allowing models (CAMs) can better resolve the physical processes responsible for heavy rainfall generation (e.g. Kain et al. 2006; Weisman et al. 2008; Duda and Gallus 2013), they too can suffer from many of these biases (e.g. Kain et al. 2006; Lean et al. 2008; Kain et al. 2008; Weisman et al. 2008; Herman and Schumacher 2016a). Furthermore, although there is a plethora of CAM guidance out to the day-ahead time frame (out to 36 hours to perhaps 48 hours after initialization), due to current computational constraints, there is almost no operational CAM guidance running out to two days ahead, and nothing operational that runs to three days ahead or beyond. Instead, global ensembles with parameterized convection serve as the primary source of forecast information and uncertainty quantification at these lead times. Even if CAMs were to integrate out to these time frames, it is not clear that the depictions would provide more valuable forecast guidance due to the rapid upscale growth of smaller-scale errors which accumulate with increasing forecast lead time (e.g. Zhang et al. 2003, 2007). Nevertheless, there is considerable utility in skillful extreme precipitation forecasts at these longer lead times, as preparative and mitigative actions may take time to execute and therefore be feasible multiple days away from the event but not just hours before. For forecasting in the medium-range beyond the day-ahead time scale, the aggregate consideration of these circumstances suggests the potential for considerable potential in applying statistical post-processing techniques to global ensemble guidance; so doing may demonstrate the capability to alleviate or even eliminate many of these dynamical model deficiencies. The increased skill at these lead times relative to even longer ones,

and operational production of extreme rainfall products at these lead times not generated even farther out in time such as the Excessive Rainfall Outlooks produced by the Weather Prediction Center (Barthold et al. 2015) further motivates a specific focus on the Day 2 and Day 3 lead times for this application.

There is a long history of successful application of statistical post-processing to dynamical model output (e.g. Klein et al. 1959; Glahn and Lowry 1972). Model Output Statistics (MOS; e.g. Glahn and Lowry 1972), is a simple, effective multivariate linear regression technique relating a set of dynamical model predictors to sensible weather predictands such as minimum and maximum temperature, wind speeds, and precipitation probability. This basic technique has long demonstrated skill over both the underlying models and even human forecasters (e.g. Jacks et al. 1990; Vislocky and Fritsch 1997; Hamill et al. 2004; Baars and Mass 2005), but is inherently limited by the linear assumptions underlying the method. Statistical post-processing techniques have also been successfully applied to QPFs, from early linear approaches (e.g. Bermowitz 1975; Antolik 2000) to more contemporary techniques that can exploit more complex variable relationships, including neural networks (e.g. Hall et al. 1999), reforecast analogs (e.g. Hamill and Whitaker 2006; Hamill et al. 2015), logistic regression (LR; e.g. Applequist et al. 2002), random forests (RF; e.g. Gagne et al. 2014; Ahijevych et al. 2016; Gagne et al. 2017), and other parametric techniques (e.g. Scheuerer and Hamill 2015). For other meteorological applications, other machine learning algorithms, such as support vector machines (e.g. Zeng and Qiao 2011; Herman and Schumacher 2016b) and boosting (e.g. Herman and Schumacher 2016b; Hong et al. 2016) have also been successfully applied. Related techniques have also been applied to forecasting related high-impact phenomena, such as severe hail (Brimelow et al. 2006; Gagne et al. 2015) and tornadoes (Alvarez 2014). One of the most powerful aspects of machine learning algorithms—and RFs in particular—is finding patterns and non-linear interactions in the supplied training data (e.g. Breiman 2001).

Depending on the extent and diversity of the data supplied in these experiments, trained RFs pose the theoretical capability of diagnosing and automatically correcting for various kinds of model biases, including context-dependent quantitative biases, such as QPF being systematically too high or too low; spatial displacement biases in the placement of extreme precipitation features; and, to some extent, temporal biases in the initiation or progression of extreme precipitation features.

This study makes a comprehensive investigation of using a global reforecast dataset to produce skillful and reliable probabilistic forecasts of locally extreme precipitation using the RF statistical post-processing technique in the medium-range. The following section provides further background and rigorously describes the data and methods used, algorithms employed, models trained, and experiments performed. Section 3 presents results of the sensitivity experiments conducted, while Section 4 presents the final results of the trained models and provides two brief case studies illustrating the process. Section 5 summarizes the findings of this study, outlines complementary analysis of these models, identifies avenues for further research, and discusses the broader implications of the results on numerical weather prediction and post-processing.

2. Data and Methods

There are several successive steps applied in creating the final forecasts evaluated in this study. A schematic overview of the forecast pipeline for the models trained in this study is depicted in Figure 1. Many types of hydrometeorological information are first taken, then assembled in a methodical manner, further pre-processed for subsequent analysis, analyzed using a statistical machine learning algorithm, and finally, extreme precipitation forecast guidance is produced and evaluated. This section details each of these setps in the model development and evaluation process.

a. Datasets

Dynamical model data used for training the RF models in this study comes from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) dataset. The GEFS/R is a global 11-member ensemble with parameterized convection and T254L42 resolution—which corresponds to an effective horizontal grid spacing of $\sim$55 km at 40° latitude— initialized once daily at 0000 UTC back to December 1984. Perturbations are applied only to the initial conditions, and are made using the ensemble transform with rescaling technique (Wei et al. 2008). The ensemble system used to generate these reforecasts is nearly static throughout its 30+ year period of coverage, though updates to the operational data assimilation system over time have resulted in some changes in the bias characteristics of its forecasts over the period of record (Hamill 2017). Some forecast fields are preserved on the native Gaussian grid ($\sim$ 0.5° spacing), while others are available only on a 1°x1° grid. Temporally, forecast fields are archived every three hours out to 72 hours past initialization, and are available every six hours beyond that. This study employs an almost 11-year period of record to explore this forecast problem, using daily initializations from January 2003 through August 2013.

In creating probabilistic extreme precipitation forecast guidance, the predictand must first be concretely specified and a robust, consistent verification framework established. One of the many challenges in heavy rainfall and flash flood forecasting is the considerable difficulty in verifying events (e.g. Welles et al. 2007; Gourley et al. 2012; Barthold et al. 2015), as every approach has its deficiencies and limitations. Flash flood reports, flash flood warnings, and quantitative precipitation estimate (QPE) exceedance of flash flood guidance (FFG)—a product issued routinely by NWS River Forecast Centers (RFCs) which estimates the amount of precipitation over a prescribed accumulation interval currently required at a given location to produce flash flooding—all attempt

to directly address the actual hydrologic impacts, but all suffer from a combination of intermittent reporting (e.g. Pielke et al. 2002), lack of coverage (e.g. Marjerison et al. 2016), and substantial regional differences in practices or methodology (e.g. Ntelekos et al. 2006; Schmidt et al. 2007; Ashley and Ashley 2008; Villarini et al. 2010; Calianno et al. 2013). It is therefore attractive to consider the problem from a simpler perspective by considering QPE exceedances of some temporally static threshold, an approach which avoids these issues while knowingly neglecting other important elements of the forecast problem. In particular, a fixed threshold (e.g. 50 mm hr^{-1}) can be used as a proxy for flash flooding (e.g. Brooks and Stensrud 2000; Hitchens et al. 2013), as can exceedances of thresholds defined relative to the local precipitation climatology (e.g. Schumacher and Johnson 2006; Stevenson and Schumacher 2014; Herman and Schumacher 2016a), such as average recurrence intervals (ARIs). An ARI defines a fixed frequency relative to the hydrometeorological climatology of the region; in particular, it corresponds to the expected duration, given the local climatology, between exceedances of a given threshold. For example, the 1-year ARI for 24-hour precipitation accumulations describes the accumulation amount for which one would expect the mean duration between exceedances of said amount to be one year. Past research has shown that a fixed-frequency ARI-based framework has better correspondence with heavy precipitation impacts than the use of any fixed threshold across the hydrometeorologically diverse regions of CONUS (e.g. Reed et al. 2007). From the perspective of forecast verification, defining extreme precipitation with respect to a fixed threshold exceedance raises challenges when applied uniformly across CONUS. Due to substantially varying frequency of event occurrence, skill differences may be artificially reflected in an undesirable manner (e.g. Hamill and Juras 2006). The ARI framework avoids this issue and provides reasonable correspondence with precipitation impacts without considering the additional influential complications such as antecedent conditions,

local hydrology, and urban effects (e.g. Herman and Schumacher 2016a). Consequently, the ARI framework is used to quantify extreme rainfall for these experiments.

Specifically, forecast probabilities are issued for 24-hour ARI exceedances at each GEFS/R archive grid point on its native Gaussian grid at all points across CONUS, using a predictand with three categories: 1) No 1-year ARI exceedance at any point within the grid point domain, 2) At least one 1-year ARI exceedance, but no 10-year ARI exceedances within the grid point domain, and 3) At least one 10-year ARI exceedance within the grid point domain. For evaluation, probabilities from the middle and most severe categories are often aggregated to produce a 1-year ARI exceedance probability. This approach has the advantage of retaining aspects of the anticipated event severity as would be retained in a regression context but is largely lost when performing single category classification. While there can be some additional complications especially with respect to calibration, formulating the prediction problem as a single multicategory classification task rather than multiple distinct binary category models also ensures mathematical consistency of the exceedance probabilities within the generated probability mass functions in a way that the latter approach would not.

In aggregating multiple QPE-to-ARI threshold grid point comparisons in a single predictand, the forecasts issued correspond to neighborhood event probabilities, which is an increasingly popular method of communicating probabilistic high-impact weather information in forecast operations (e.g. Barthold et al. 2015; NWS 2017a). While the event frequency does increase relative to the purported ARI, the fixed-frequency property and associatedly many of the aforementioned desirable properties of the framework are approximately retained. For this study, focus is placed exclusively on two 24-hour forecast periods: the 1200–1200 UTC period corresponding to forecast hours 36–60 from the GEFS/R forecast fields and the subsequent 24-hour period encompassing forecast hours 60–84, denoted respectively as Day Two and Day Three. At these times, there is

typically some knowledge to characterize the environmental conditions in which precipitation may form, but it is beyond the current range of operational CAM guidance.

The National Centers for Environmental Prediction (NCEP) Stage IV Precipitation Analysis (Lin and Mitchell 2005) is an operational QPE product created daily since December 2001. Stage IV provides 24-hour analyses over the CONUS on a $\sim$4.75 km grid. It uses both rain gauge observations and radar-derived rainfall estimates to generate an analysis, and is further quality controlled via NWS River Forecast Centers (RFCs) to ensure stray radar artifacts and other spurious anomalies do not appear in the final product. Despite some limitations (Herman and Schumacher 2016a; Nelson et al. 2016), its analysis quality; resolution, allowing better ability to capture precipitation extremes compared with other QPE products (e.g. Hou et al. 2014); and data record length make it preferable to other precipitation analysis products, and is therefore be used as the precipitation 'truth' for verification purposes in this study.

The ARI thresholds associated with the 1- and 10-year ARIs for 24-hour precipitation accumulations are generated using the same methodology of Herman and Schumacher (2016a), where CONUS-wide thresholds are produced by stitching thresholds from several sources. NOAA's Atlas 14 thresholds (Bonnin et al. 2004, 2006; Perica et al. 2011, 2013), an update from older work and currently under development, are used wherever were available at the commencement of this study. For five northwestern states—Washington, Oregon, Idaho, Montana, and Wyoming—updated thresholds are not available, and derived NOAA Atlas 2 threshold estimates are used instead (Miller et al. 1973). Additionally, in Texas and the Northeast—New York, Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island—Technical Paper 40 (TP-40; Hershfield 1961) thresholds are used[1]; everywhere else uses the Atlas 14 threshold estimates.

[1]The northeastern states did receive updated Atlas 14 estimates in October 2015, but TP-40 thresholds were retained for consistency with prior work.

These 1- and 10-year thresholds are depicted in Figure 2 in panels (a) and (b), respectively. The 10-year ARI thresholds in Figure 2b show a similar pattern to the 1-year ARI thresholds in the same location shown in Figure 2a, but are substantially higher everywhere. More significantly, it is apparent that at both severity levels, there are large regional disparities in the threshold magnitudes. Over climatologically wet regions of CONUS, such as the Pacific coastal mountains and immediately along the Gulf Coast, thresholds are as high as 100–150 mm and 250–300 mm for 1-year and 10-year ARIs, respectively. Thresholds, particularly over central and eastern CONUS, tend to decrease smoothly with increasing latitude and distance from major bodies of water. Sharper variations are seen in areas of complex terrain over western CONUS. In the driest parts of the arid southwest and intermountain west, thresholds can be as low as 10–15 mm and 25–30 mm for the two ARI levels—a full order of magnitude difference from the largest thresholds at the same intensity level. This highlights the stark contrast the use of the ARI framework has with the use of a fixed threshold framework across all of CONUS.

Forecast models in this study are trained separately for eight distinct, yet cohesive and internally fairly hydrometeorologically homogeneous regions of CONUS, using the delineation indicated in Figure 3. Observed 1- and 10-year ARI exceedance events that occurred during the period of record are depicted by region respectively in Figures 2 (c) and (d). There are important regional differences in the seasonal cycles and climatology of ARI exceedances across CONUS. In the Pacific Coast (PCST) region, the vast majority of exceedances at both the 1-year and 10-year severity levels occur in the cool-season, and occur largely from atmospheric river events with large moisture transport impinging on coastal topography (e.g. Rutz et al. 2014; Herman and Schumacher 2016a). This seasonality holds to a lesser extent in the neighboring Southwest (SW) region, with some signal carrying over to the Rockies (ROCK) region as well. In the central and eastern regions, the majority of events occur during the warm-season from more scattered convective-scale

processes, particularly in the months of May, June, and July (e.g. Schumacher and Johnson 2006; Herman and Schumacher 2016a). Tropical cyclones can cause widespread and very significant rainfall, and comprise a substantial portion of the extreme precipitation climatology, especially in the Northeast (NE) and Southeast (SE) regions. Due to the spatial extent of their impacts and immense rainfall totals they can produce, they form a much larger fraction of the climatology of 10-year ARI exceedances (Fig. 3d) than 1-year events (Fig. 3c). Additionally, the numbers are lower than would be expected; by the explicit exceedance frequencies associated with the thresholds, one would expect an average of one exceedance per point per year over the period of record for the 1-year events (Fig. 3c) and 0.1 exceedances for 10-year events (Fig. 3d); in reality, event counts are only about half of that. This is consistent with previous findings (e.g. Herman and Schumacher 2016a), and likely in part attributable to limitations in the Stage IV product to capture extremes (e.g. Nelson et al. 2016). There is also quite a bit of region-to-region variability in event counts, particularly for 10-year exceedances, much of which is attributable to statistical variability from having a short data record in relation to the event frequency.

b. Predictor Assembly

Input predictors, or features, to the random forests can be partitioned into two categories: model predictors and background predictors. Model predictors, which constitute the vast majority of the total number of inputs, come from atmospheric fields forecast in the GEFS/R which bear a known physical relationship with extreme precipitation. A core set of f=9 fields used in this study are: accumulated precipitation (APCP), convective available potential energy (CAPE), convective inhibition (CIN), precipitable water (PWAT), surface temperature (T2M) and specific humidity (Q2M), surface zonal (U10) and meridional winds (V10), and mean sea level pressure (MSLP). Sensitivity experiments explore the use of additional upper-air atmospheric fields; a full list of

fields used in this study, their associated symbols used in this manuscript, and the grids on which they are each archived are tabulated in Table 1. In addition, considering different atmospheric fields in forecasting extreme precipitation, the spatiotemporal variations in these fields are considered as well. Spatially, predictors are structured in a forecast-point relative sense. In the control model, GEFS/R forecast values up to r=4 grid boxes ($\sim 2^\circ$) latitudinally or longitudinally displaced in any direction relative to the forecast point are considered. Temporally, portrayed fields are considered at each archived time during the forecast interval, which corresponds to every three hours during the Day Two period and every six hours during the Day Three period, for a total of t=9 and t=5 forecast periods for the Day 2 and 3 periods, respectively. All told, this yields $tf(2r+1)^2$ model predictors, which yields respectively M=6,561 and M=3,645 model predictors for the Day Two and Day Three control models. The other category of predictors, background predictors (Table 2), are those which are solely associated with the forecast point, and have no relation to the present meteorology. These include the location of the point, as well as the ARI characteristics of the point and in the surrounding area.

c. Dimensionality Reduction

There are a large number of model predictors, and they are also highly correlated—spatially, temporally, and across variables. With millions of training examples and thousands or, in some cases, tens of thousands of features, the forecast problem can become computationally intractable. Further, having many highly correlated features can readily result in model overfitting—making predictions based on noise affecting an individual native feature rather than the underlying signal—a phenomenon commonly termed the "curse of dimensionality" (e.g. Friedman 1997). There are numerous ways these concerns can be addressed; broadly speaking, the most common approaches are either feature selection or feature extraction. In feature selection, a subset of initial predictors

are chosen that collectively bear the strongest predictive relationship with the predictand, whereas in feature extraction, a smaller set of new predictors are derived from the original set. Both of these procedures can be performed subjectively through manual means or objectively through automated means. In this case, all of the input predictors are believed to have a physical relationship with extreme precipitation, and choosing only the most predictive fields (e.g. model QPF) and discarding the rest risks removing valuable predictive information not contained in the retained predictor set. The primary issue with the input predictors in this case is not that many may not have any physical bearing on the predictand, but rather that each predictor represents a value at a different point of a continuous field, or a different property at the same point, and are thus necessarily highly correlated to one another. Furthermore, while one could conceivably extract features using field averages or some other pre-determined method, this may not be optimal. For example, it may be better to weight values closer to the forecast point more heavily, while still retaining some information from the far-field predictors. Holistically assessed, this suggests that an objective approach to feature extraction rates to be the most beneficial approach to dimensionality reduction. Though it has some limitations (e.g. Shlens 2014), principal components analysis (PCA; Ross et al. 2008; Pedregosa et al. 2011) is a robust and frequently utilized approach for dimensionality reduction by feature extraction. This creates a small set of uncorrelated predictors that explain the signal in the forecast data and gives insight into the regional modes of atmospheric variability as depicted in the GEFS/R model (explored in more depth in a companion paper), while leaving the noise in lower-order principal components (PCs), acting in principle to both alleviate overfitting and manage computational requirements.

d. Machine Learning Algorithms and Sensitivity Experiments

The primary statistical algorithm used in this study is random forests (RFs; Breiman 2001). RFs are in essence an ensemble of *decision trees*, where traditionally each tree individually makes a deterministic prediction about the outcome of the predictand; the relative frequencies of each possible predictand outcome in the ensemble of trees are then used to make a probabilistic forecast. Much further detail on tree and RF construction and mechanics can be found in Appendix A. There are also several parameters which can be tuned to the particular forecast problem in order to maximize model performance. Four-fold cross-validation is used for model development in this study, whereby each model configuration examined is trained four times, once each on three-quarters of the training data, and then evaluated on the final withheld quarter. Each quarter of the training data is a temporally contiguous chunk to avoid issues of sample independence and approximately mimic information that would be available in an operational context. All parameter settings and sensitivity experiments are evaluated in this framework. The set of RF parameters tuned is described in Appendix A, and the results presented in Appendix C.

In this study, there are a great deal of dynamical model data considered as input information on which the RF can base a prediction. An important question in the design of statistical post-processing algorithms in a resource-constrained environment is which of this information actually improves forecasts, and the relationship between extent of additional dynamical model information and statistical model skill. In order to obtain answers to these important questions, a suite of sensitivity experiments are conducted as summarized in Table 3. Sensitivity to the inclusion of horizontal variations in atmospheric fields is explored by varying the previously described predictor radius parameter R from 0 to 4. The effect of including predictor information from additional upper-air fields is explored by comparing the inclusion and exclusion of two sets of fields. The first

incorporates temperature, specific humidity, zonal and meridional winds at 850 and 500 hPa, and 850 hPa vertical velocity in the so-called Upper-Air Core predictor group, while further inclusion of those same fields at 700 and 250 hPa—the so-called Upper-Air Extra predictor group—allows a reasonably comprehensive view into the value of this forecast information. Investigation into the utility of temporal resolution of simulated atmospheric fields in generating skillful forecasts is also performed. Predictor density is three-hourly for Day 2 guidance and six-hourly for Day 3 guidance; models are additionally trained with predictors at twelve-hourly temporal density for both lead times and six-hourly temporal density for the Day 2 forecast model to ascertain this aspect. Another important element—and one with implications for how operational centers allocate their computational resources—is the extent of use of ensemble information and its implication on RF model skill. Using forecast information from only the GEFS/R's control member in model training is compared with using the ensemble median from the full ensemble, and then further with the use of the ensemble second-lowest and second-highest values for each atmospheric field in conjunction with the median to evaluate the impact of this dimension of forecast information, following the findings of Herman and Schumacher (2016b), which found relatively little sensitivity in performance with respect to how ensemble information is used, but using the near-minimum, median, and near-maximum values outperformed using the mean and spread. Additionally, where possible, models are trained with and without the aforementioned PCA pre-processing step, and an assessment of the effect of this pre-processing step on model skill is made by comparing the two. An additional sensitivity experiment explores the effect of region size on forecast skill, hypothesizing models trained for larger regions may exhibit higher skill due to more available training data. This is performed by aggregating the ROCK and SW regions into a new WEST one, combining the Southern Great Plains (SGP), Northern Great Plains (NGP), and Midwest (MDWST) regions into a CENTRAL region, and collecting SE and NE regions into a single EAST region, while

leaving PCST, with its highly unique extreme precipitation climatology unperturbed. A final sensitivity experiment investigates model performance as a function of model algorithm, specifically by comparing with logistic regression (LR), a common and comparatively simpler alternative to statistically deriving forecast probabilities. Further discussion of LR and other machine learning alternatives to the RF algorithm is included in Appendix B.

e. Model Evaluation

Based on the parameter tuning and sensitivity experiment results, final model configurations are selected. The final model is run over a completely withheld 4-year evaluation period spanning September 2013–August 2017. The forecasts gerenated from the final model are compared with those from the full ensemble of raw GEFS/R QPFs, as well as the full 50-member ECMWF global ensemble, accessed from TIGGE (Molteni et al. 1996; Bougeault et al. 2010). The comparison with the former provides an assessment of what improvement, if any, these models yield compared with the raw guidance from which their forecasts are derived when evaluated in a real-time setting. The latter, meanwhile, provides an assessment for how these forecasts compare with state-of-the-science operational ensemble guidance available at these lead times. To make these comparisons, the QPF from each ensemble member of the two ensembles is regridded onto the $\sim$4.75km Stage IV HRAP grid on which the Atlas thresholds lie using a first-order conservative scheme (Ramshaw 1985). These regridded QPFs are then compared with the 1-year and 10-year ARI thresholds to create deterministic exceedance forecasts with respect to the two thresholds for each ensemble member. These binary grids are then upscaled to the GEFS/R grid using the same procedure as the verification upscaling: any exceedance in the downscaled grid corresponds to an exceedance at the nearest GEFS/R point in the upscaled grid. Since the predictand categories are necessarily mutually exclusive, the 1-year ARI exceedance grids are modified so that any member forecasting

a 10-year ARI exceedance at a point is not forecasting a between 1-and-10-year exceedance at that same point and time period. The prevailing operational method of generating forecast probabilities from a dynamical ensemble—democratic voting, whereby the fraction of ensemble members forecasting the event is used as the forecast probability (e.g. Buizza et al. 1999; Eckel 2003)—is applied to each ensemble to generate the exceedance probabilities for the reference forecasts.

Skill, both in the final assessment of model performance as well as in all aforementioned sensitivity experiments, is quantified by means of the Rank Probability Skill Score (RPSS) with a climatological reference:

$$RPSS = 1.0 - \frac{\Sigma_{d=1}^{D}(\Sigma_{p=1}^{P}(\Sigma_{m=1}^{K}(\Sigma_{j=1}^{m}P_{jpd} - O_{jpd})^2))}{\Sigma_{d=1}^{D}(\Sigma_{p=1}^{P}(\Sigma_{m=1}^{K}(\Sigma_{j=1}^{m}P_{clim_j} - O_{jpd})^2))} \tag{1}$$

with D forecast days; P forecast points; K predictand categories; P_{jpd} and O_{jpd} corresponding respectively to the forecast probability and observance of predictand category j on day d and point p; and P_{clim} corresponding to the climatological frequency of occurrence, as defined by the respective ARIs of the predictand. A score of 1.0 indicates a perfect forecast, and a score of 0.0 indicates model performance equivalent to forecasting climatology. Final assessment also includes analysis of reliability, both subjectively through reliability diagrams, and quantitatively via the Murphy (1973) decomposition of the Brier score (BS) for category j^*:

$$BS_{j^*} = \Sigma_{n=1}^{N}(P_{Nj^*} - O_{Nj^*})^2 = \frac{1}{N}\Sigma_{c=1}^{C}N_{cj^*}(P_{cj^*} - \overline{O_{cj^*}})^2 - \frac{1}{N}\Sigma_{c=1}^{C}N_{cj^*}(\overline{O_{j^*}} - \overline{O_{cj^*}})^2 + \overline{O_{j^*}}(1 - \overline{O_{j^*}}) \tag{2}$$

where there are $N = DP$ total forecasts, broken into C discrete probability bins with N_c forecasts being issued for each bin c. $\overline{O_{j^*}}$ denotes the climatological (based on the period of record) frequency of observing event category j^* and $\overline{O_{cj^*}}$ denotes the proportion of forecasts in probability bin c observing event category j^*, where j^* is the aggregation of event categories of at least j in the RPSS framework. $\overline{O_{j^*}}(1 - \overline{O_{j^*}})$, the so-called "uncertainty" term, also represents the BS of a

climatological forecast. Converting to a Brier skill score (BSS) framework by dividing out by this term:

$$BSS_{j^*} = 1.0 - \frac{BS_{j^*}}{BS_{clim_{j^*}}} = \underbrace{\frac{\frac{1}{N}\sum_{c=1}^{C} N_{cj^*}(\overline{O_{j^*}} - \overline{O_{cj^*}})^2}{\overline{O_{j^*}}(1-\overline{O_{j^*}})}}_{\text{“Resolution”}} - \underbrace{\frac{\frac{1}{N}\sum_{c=1}^{C} N_{cj^*}(P_{cj^*} - \overline{O_{cj^*}})^2}{\overline{O_{j^*}}(1-\overline{O_{j^*}})}}_{\text{“Reliability”}} \quad (3)$$

This analysis is conducted for both the 1- and 10-year thresholds.

3. Results: Sensitivity Experiments

Examining forecast skill as a function of time step between atmospheric field predictors (i.e. the CORE_LTIME models of Table 3; Fig. 4a), two striking findings concern 1) the large variations in forecast skill across regions and 2) the evidently low sensitivity of time step length and forecast skill within any given region. For the 3-hour time step, predictors are gathered from a total of 9 forecast times; with the 6-hour step, 5 forecast times are used; and with the 12-hour time step, a total of 3 forecast times are used. The 12-hour time step therefore has one-third the total number of predictors as the model with the 3-hour time step, but still yields nearly identical forecast skill results. In most regions and forecast periods, there is a slight degradation in performance going from the 6- to 12-hour time step, but the difference is not generally statistically significant by a 99% bootstrap skill score difference test (not shown).The one exception to this is in the PCST region, which has much higher skill overall than the other regions for both forecast periods, and exhibits somewhat higher sensitivity to the predictor time step than the other regions, particularly in going from 6-hours to 12-hours, with skill differences of approximately 0.01.

Similar to the temporal resolution findings, a general lack of sensitivity as a function of predictor spatial extent (Fig. 4b). This finding comes in stark contrast to that of Herman and Schumacher (2016b), which found great sensitivity of predictor spatial extent in forecasting airport flight rule conditions. Albeit weak, a slight improvement in skill for most forecast period, region combi-

nations can be noted with increasing predictor radius, often to the extent that the skill difference between 0 and 4 grid box radii is statistically significant (not shown). Two regions in particular, the NE and PCST, exhibit by far the most sensitivity to predictor spatial extent, with differences of roughly 0.02 observed over the evaluated interval. Also of note is that a radius of 4 grid boxes—the highest number evaluated—did not always yield the best performance results; most notably, the Day 2 model for the NE region maximized skill at a radius of 2, with a slight deterioration of forecast skill with increasing radius thereafter. In those regions where the GEFS/R cannot explicitly resolve the processes responsible for producing extreme precipitation, the RF is ultimately making forecasts more on environmental factors; these do not vary drastically in time or space, and thus a single number or small set of numbers at or immediately surrounding the forecast point are sufficient to characterize the basic properties of the environment, and this is all that the RF is really using for much of its predictions. However, in regions impacted more readily by larger scale systems where the dynamical model can more directly simulate the precipitation processes such as PCST and the NE, the spatial variations in atmospheric fields carry more signal rather than noise and thus contribute more predictive value.

Like varying spatial and temporal density, there is relatively little sensitivity to the inclusion of more atmospheric fields (Fig. 5a). Slight but consistent improvement is observed in adding the core upper-air fields as predictors, but adding further levels beyond the core group was found to not improve predictive skill, and actually resulted in a decrease in skill for the PCST, NE, ROCK, and SE regions—those which are most affected by larger scale precipitation systems. Though still rather small, somewhat more distinct sensitivity to type of ensemble information included (Fig. 5b) can be seen here across all regions, with improvements seen using only predictor information from the GEFS/R the ensemble median versus the control member, and slight additional improvement using the ensemble second-from-minimum and second-from-maximum in addition to the

ensemble median. The largest differences in magnitude are again for the PCST region, but in this experiment, clear and statistically significant differences (not shown) are also seen for low skill, convectively active regions such as MDWST.

Aggregating regions (Fig. 5c) results in a slight degradation in forecast skill. While predictor information is available to pinpoint the locations via latitude and longitude and then perform further analysis from there, effectively performing the manual regional partitioning in the CTL_NPCA model, these findings demonstrate that there is some—albeit limited—utility in manually partitioning training data with distinct hydrometeorological relationships, rather than relying on the machine learning algorithm to discern the distinction automatically. Comparing the impact of applying PCA pre-processing to the RF (Fig. 5d, leftmost two columns), performing PCA tends to either improve performance, as is the case for the PCST, NE, SW, and MDWST regions, or make little difference, as seen in the ROCK, NGP, SGP, and SE regions. The positive differences tend to be larger in magnitude, both in relative and absolute senses, for Day 2 model versions compared with Day 3. Forecasts produced through LR tend to be substantially worse than those generated by RFs (Fig. 5d, center columns). However, the exact magnitude to which this is the case varies by region; substantial differences in skill are seen between RF and LR forecasts for the SW, ROCK, and SGP regions, while there is almost no skill difference between the Day 3 forecasts in the PCST region. This may suggest the linear assumptions inherent to the LR algorithm perform better in larger scale systems than in the more convectively active ones in which the responsible processes are highly nonlinear, but this causality is not entirely clear. Finally, a weighted average of RF and LR forecasts outperforms its component members for all regions and forecast periods. The extent of overperformance is strongly tied to the skill difference between the RF and LR models; when the skill difference is small, the value of the weighted average is comparatively large to when the RF performs much better than LR (cf. Fig. 5d PCST and SW lines). Since these weighted av-

erages performed the best in cross-validation, a weighted average using each of the CTL_NPCA, CTL_PCA, and CTL_LR models was chosen for the final model configuration.

4. Results: Final Model Performance

For both the final ML models and the forecasts from the raw QPFs of both the GEFS/R and ECMWF (Fig. 6), a usually statistically significant[2] deterioration in forecast skill from Day 2 to Day 3 is evident in each CONUS region over the four year test period. Forecast skill is significantly higher in regions with extreme precipitation associated partially or primarily with synoptic scale precipitation episodes, such as PCST, SW, and ROCK, rather than smaller scale convective systems that characterize extreme precipitation in the NGP, SGP, and MDWST regions. At an extreme, the NGP and SGP GEFS/R raw QPFs have no skill in predicting 1-year and 10-year ARI exceedances. Especially for the ML models, the bigger Day 2 vs. Day 3 skill differences are also seen where the skill is higher, again suggesting the direct forecasting of the precipitation as opposed to forecasts more reflecting the forecast environment, either dynamically via parameterized convection in the case of raw QPFs, or directly in the case of the ML model forecasts. Furthermore, the ML models exhibit a larger skill deterioration between Days 2 and 3 than either of the raw ensemble forecast sets.

[2]For all of the comparisons in this section, statistical significance is assessed by bootstrapping to obtain identical sets of cases for each of the two forecast sets being compared. Skill scores are derived from the subsample of each forecast set, and a skill difference is computed. This process is repeated 1000 times to generate a distribution of skill differences, and statistical significance is ascertained with respect to whether the 0.5^{th} and 99.5^{th} percentile skill score difference values from the bootstrap trials overlap zero. This 99% confidence bound is used in contrast to 90% or 95% bounds to compensate for concerns arising from conducting statistical significance analysis on numerous different comparisons. While some uncertainty analysis has been included in the figures, much of the statistical significance difference results discussed in-text are omitted for the sake of concision.

Comparing the forecast systems, the ECMWF forecasts consistently and statistically significantly outperform the GEFS/R forecasts at all lead times except in the SE region (Fig. 6). Encouragingly, the ML model forecasts are statistically significantly more skillful for all eight regions and both lead times compared with the GEFS/R forecasts from which they are based. The post-processing is thus clearly accomplishing its purpose of improving forecast skill. But it is also apparent that the GEFS/R is not a state of the science model for extreme QPF prediction given its lower skill compared with the ECMWF. The real test of the ML model then is how it compares with current best operational guidance for these lead times, represented here with the ECMWF ensemble. The comparison (Fig. 6) is generally quite favorable, with the Day 3 ML forecasts outperforming the Day 2 ECMWF forecasts across all regions except ROCK and PCST. In the non-western regions, the extent of overperformance is quite considerable when comparing equal lead times, with skill score improvements of factors of two to three seen in many comparisons. In the ROCK and PCST regions, the ML and ECMWF forecasts performed about equally at Day 2, and ECMWF performed slightly better at Day 3. Overall though, the ML models demonstrated ability to consistently outperform current operational model guidance, especially in convectively active regions where there is no operational guidance that can dynamically resolve the physical processes producing extreme precipitation.

Reliability diagrams of raw GEFS/R forecasts (Fig. 7) reveal highly overconfident probabilistic exceedance forecasts for all regions, lead times, and severity levels as evidenced by the shallow slope relative to the one-to-one line. The raw GEFS/R forecasts are relatively sharp, with more than 0.01% of forecasts falling into each probability bin above 10%, and a vast majority of zero probability forecasts (not shown). For all regions, there are cases where every ensemble member has simultaneously predicted a 1-year exceedance (Fig. 7a,c), but the same is not true for 10-year exceedance predictions in the northeastern regions: NE, NGP, and MDWST (Fig. 7b,d).

The ECMWF (Fig. 8) is also overconfident, but we see that it is also negatively biased for all regions, lead times, and severity levels. Its degree of overconfidence is dampened compared with the GEFS/R, and it is not as sharp, with fewer very occurrences of very high forecast probabilities except in the westernmost regions of ROCK and PCST (Fig. 8a,c, inset panels). With 50 members rather than 11, there is also substantially more resolution across the probability spectrum in the ECMWF forecasts. By the very nature of how these forecasts are generated, quite a bit of sharpness is inherent at the cost of reliability, since it is not possible for probabilities near the climatological event frequency to be issued for either raw ensemble, particularly for the GEFS/R.

The reliability diagrams for the different components of the final model, the CTL_NPCA, CTL_PCA, and CTL_LR models are shown respectively in Figures 9, 10, and 11. The CTL_NPCA (Fig. 9) shows markedly different characteristics than either of the raw ensembles. In particular, all of the regions exhibit an underconfidence signal, with low probability events below about 2% for 1-year events (Fig. 9a,c) occurring with observed relative frequencies below the forecast probabilities. The relative event frequencies are conversely appreciably higher than the forecast probabilities would indicate for probabilities above 5%. The PCST probabilities are the most negatively biased, while NE probabilities are the most positively biased, among the regions. Overall, reliability is much better than for either raw ensemble, but this comes at the expense of sharpness. Less than 1 in 10,000 forecasts are above about 20% for 1-year exceedance probabilities for any CTL_NPCA model (Fig. 9a,c, inset panels), or 5% for 10-year events (Fig. 9b,d, inset panels), and maximum probabilities are in the 30–80% range for 1-year exceedance forecasts depending on the lead time and region, compared with 100% for all lead times and regions in the raw ensembles. For the 10-year events, maximum probabilities range from 10%–40%, again much lower than the maximum probabilities from the raw ensembles, which were still at or near unity (e.g. Fig. 7b, 8b).

The CTL_PCA model (Fig. 10) exhibits very similar reliability characteristics to the CTL_NPCA model (cf. Fig. 9, 10), including the underconfidence, reduced sharpness compared with the raw GEFS/R, and different regional probability bias characteristics. The probability distribution for 1-year exceedance events is not markedly different between the Day 2 and Day 3 forecasts (cf. Fig. 10a,c), but the relatively higher probabilities issued for 10-year exceedances in Day 2 do not occur at the Day 3 lead times (cf. Fig. 10b,d). This is consistent with increasing confidence in very extreme events with decreasing lead time; something seen very pronounced in the CTL_PCA model and to a slightly lesser extent in the CTL_NPCA model, but to a much lesser extent in the raw ensemble forecasts. The CTL_LR model (Fig. 11) exhibits some similarities and some differences with the RF-based models. PCST forecasts are consistently the most negatively biased, followed by ROCK and the SE, with NE region forecasts being the least negatively biased. However, unlike the RF-based forecasts, the LR model issues more high probabilities; at Day 2, for example (Fig. 11a), forecasts in the highest probability bin were issued for most regions. At the highest probabilities, the forecasts revert to being positively biased, as they are for events with probabilities issued in the 0.01–1% range. At very low probabilities (not shown), LR-based forecasts are substantially more negatively biased than for RF-based forecasts, leading to considerable overconfidence overall when considering that the vast majority of forecasts issued occur on this low probability end of the spectrum. While LR (and regression in general) is effective at removing bias in a global sense, since a single regression equation must necessarily apply globally to all forecasts, it inherently cannot perform more localized, context-depend forms of bias correction, leading to forecast probability-dependent model biases. Finally, the final ML model reliability (Fig. 12) unsurprisingly reflects a blend of the component members, retaining some of the underconfidence of the RF-based models while adding a bit of sharpness from the CTL_LR model in

regions where it verified skillfully enough in cross-validation (e.g. PCST, Fig. 5d) to garner much weight.

The relationship between the reliability analysis and skill via the Brier score decomposition (Murphy 1973) quantitatively solidifies many of the general observations discerned by inspection of the reliability diagrams. Though sharper than competing forecasts, the raw GEFS/R forecasts consistently exhibit the worst resolution component contribution to forecast skill for all regions and severity levels, both for Day 2 forecasts (Fig. 13a,c) and Day 3 forecasts (Fig. 14a,c) due to an ability to actually distinguish events from non-events by resolving the responsible physical mechanisms. The final ML models exhibit better resolution term skill contributions than the ECMWF ensemble forecasts, with the exception of the ROCK and NGP regions for 1-year events (Fig. 13a, 14a). Between the component models, resolution term skill tended to best for CTL_NPCA forecasts over the test period, particularly at the 10-year severity level (e.g. Fig. 13c) but the extent of the difference tended to be relatively small and there were numerous instances where PCA-based models exhibited more resolution. The weighted average consistently exhibited higher resolution than any of the component members. With respect to the reliability contribution to skill (Fig. 13b,d Day 2; Fig. 14b,d Day 3), ECMWF forecasts were, perhaps surprisingly given the lack of explicit calibration, the most reliable forecast set for all regions and lead times, while in many cases the ML models had a more negative contribution to the total skill than the raw GEFS/R, likely resulting from the underconfidence. The resolution term is at largest one and at least zero in this decomposition, while the reliability term is at most zero. The magnitude of the resolution terms is consistently several factors larger than the reliability term for all forecast sets, and the differences in that term are generally of more absolute impact on the overall Brier skill scores.

Lastly, while by no means a comprehensive characterization of the system, a sample of real cases over the test period are presented to illustrate some of the strengths and weaknesses of the

system. On the evening of 19 May 2015 and morning of 20 May 2015, a vigorous mesoscale convective system developed over southern Oklahoma and northern Texas, producing very heavy rainfall that contributed to historic flooding in the region during May 2015 (e.g. Wolter et al. 2016). Stage IV analysis (Fig. 15a) reveals that the 24-hour precipitation totals exceeded 1-year ARI thresholds within much of an E/W band encompassing the region, with embedded areas of 10-year exceedances along the state border region. While the ECMWF ensemble forecasts indicate some possibility of extreme precipitation in that region during this time frame three days out (Fig. 15c), the probabilities are displaced two far to the south and west, and the probabilities of 10-year exceedances are very low. There is some improvement in positioning with the Day 2 forecast (Fig. 15b), but remains too far west, with probabilities still quite low particularly at the 10-year ARI level. Raw GEFS/R forecasts at Day 3 (Fig. 15e) indicate quite high risk for a 1-year exceedance over a fairly narrow area, better positioned than the ECMWF ensemble at the same lead time but still too far to the west. Outside of this area, the GEFS/R indicates almost no risk of an extreme rainfall event, and also indicates no risk of a 10-year exceedance anywhere in the domain. The Day 2 forecast (Fig. 15d) looks similar to the Day 3 outlook, except that the probabilities are reduced somewhat in the target area, which also has incorrectly displaced further to the south and west. The ML model depicts a much different picture. It exudes much less confidence, with lower maximum probabilities compared with either raw ensemble, but non-zero exceedance probabilities of both 1- and 10-year exceedances across much of the domain for both Days 3 (Fig. 15g) and Day 2 (Fig. 15f). Importantly, the model elevated probabilities compared with the raw guidance in the place that extreme precipitation was actually observed (to the east of where it was forecast in the GEFS/R). In fact, at Day 2 (Fig. 15f), the probability maximum is located right where the heaviest precipitation actually occurred, displaced well to the north and east of where it was forecast in the GEFS/R (Fig. 15d). Additionally, while still low, the 10-year event probabilities are correctly

much higher over the verifying area when compared with either raw ensemble, with maximum Day 2 probabilities of around 30% and 3% for 1-year and 10-year exceedances, respectively. Finally, in contrast to the raw guidance, the ML model became increasingly confident in an event occurring with decreasing lead time (cf. Fig. 15f,g).

A different mesoscale precipitation produced extreme precipitation over southwestern Wisconsin, southeastern Minnesota, and northeastern Iowa during the evening hours and overnight hours of 21 September and 22 September 2016, respectively. Much of the area experienced 1-year ARI exceedances for the 24-hour period ending 1200 UTC 22 September 2016, and within the 1-year exceedance area, there were many embedded cells that produced 10-year ARI exceedances (Fig. 16a). ECMWF forecasts at Day 3 indicated risk of extreme rainfall, even at the 10-year severity level (Fig. 16c), but the location was poor, with exceedance probabilities high in eastern Minnesota and northern Wisconsin where extreme rainfall was not observed, and very low probabilities in northeastern Iowa and southeastern Wisconsin where it was. Both the positioning and risk of very extreme precipitation improved for the Day 2 forecst issuance (Fig. 16b), but probabilities still remained too far to the north. The GEFS/R at Day 3 (Fig. 16e) indicated very little risk of extreme precipitation in the area, with just one member correctly predicting a 1-year exceedance in southeastern Minnesota. The risk of an event occurring within the domain increased for the Day 2 issuance, but the locations got worse, with maximum risk indicated in eastern Nebraska, western Iowa, and northeastern Wisconsin, with the only 10-year prediction occurring in the latter location. Somewhat like the raw GEFS/R, the ML model had only some indication of extreme precipitation risk at Day 3 (Fig. 16g). However, it both has the higher probabilities (near 10% in both cases) distributed over a much larger area, and indicates some risk of a 10-year event, with probability maxima near 1.5%. Additionally, it has the maximum probability axis nearly collocated with where heaveist precipitation occurred, well to the south of the ECMWF probabilities, albeit still

slightly too far to the north. The Day 2 forecast issuance (Fig. 16f) is largely similar. The two main changes are a correctly increased risk in the area where the event actually verified, and an incorrectly increased risk of heavy precipitation in eastern Nebraska where the raw GEFS/R had heavy precipitation on Day 2 (Fig. 16d).

5. Discussion and Conclusions

An ML model based on RFs and LR is used to generate CONUS-wide probabilistic forecasts for the exceedance of 1- and 10-year ARI thresholds for 24-hour precipitation accumulations during the Day 2 and Day 3 periods. Approximately eleven years of GEFS/R forecasts, in particular the ensemble median, are used to train these models, and forecasts are using simulated atmospheric fields and varying in both space and time (Table 1), in addition to a variety of geographic and climatological forecast predictors (Table 2). Separate models are trained for each of the two 24-hour periods and for each of eight different regions of CONUS, as depicted in Figure 3. A variety of sensitivity experiments are performed, as outlined in Table 3, to ascertain the utility of different aspects of forecast information in predicting locally extreme precipitation. Finally, the final forecast models were evaluated, and compared with forecasts based only on the ensemble of raw QPFs from the GEFS/R and ECMWF. The ML models trained in this study demonstrably outperformed the raw GEFS/R forecasts for all regions and forecast lead times (Fig. 6), often more than doubling the forecast skill and adding substantially more than 24-hours lead time improvement in forecast skill. With the exception of the PCST and ROCK regions, the same held for comparison of the ML model forecasts with ECMWF ensemble forecasts as well. Both raw ensembles tended to be negatively biased and overconfident in predicting extreme QPFs (Fig. 7,8), particularly at the 10-year ARI for central CONUS regions; this was reversed in the final ML model forecasts, which were more reliable at higher probabilities, but generally underconfident (Fig. 12).

In general, unlike past studies (e.g. Herman and Schumacher 2016b), in most regions, the temporal resolution and extent of spatially displaced predictors from the forecast point considered had little to no impact on forecast skill (Fig. 4), in addition to the use of upper-level information and additional ensemble information (Fig. 5). These results are suggestive of two findings. First, most of the relevant information about predictors displaced spatiotemporally from the forecast point, other atmospheric fields, or other ensemble member information, can be derived with at least moderate accuracy using just the information from the ensemble median from a group of core set of fields collocated and concurrent with the forecast; that is, these additional predictors contain only limited independent forecast information, at least for this coarse dynamical model and this underdispersive ensemble configuration. It also suggests that, for the most part, the predictive ability is coming primarily through a characterization of the overall environment, which can be reasonably summarized with only a subset of predictors, rather than the simulated spatiotemporal variability and full 3-D characterization of the atmospheric evolution in the underlying dynamical model. This finding comes in contrast to similar studies of other forecast problems using the GEFS/R, such as the Herman and Schumacher (2016b) study which investigated using the GEFS/R to create ML-based probabilistic forecasts of cloud ceiling and visibility at different airports and found considerable value in the inclusion of spatially displaced predictors. However, there is at least one major exception; none of this really held for the PCST region; here, more complex models with more predictors did notably improve forecast skill. This is perhaps in part because the physical processes associated with extreme precipitation are much better resolved in the GEFS/R in this region compared with the others, and so the added information adds usable forecast utility beyond simply duplicatively characterizing the atmospheric environment for the forecast. The largest skill difference of the sensitivity experiments came for most regions in changing algorithmic assump-

tions and processes (Fig. 5d); the simpler linear assumptions of LR tended to degrade forecast skill compared with the more limited assumptions underlying the RF models.

The results of this study reveal that the application of more sophisticated statistical methods and ML algorithms such as RFs can demonstrably improve forecasts of extreme precipitation and potentially other rare, high-impact weather events in the medium range when compared with the methods and techniques that are most prevalent in forecast operations today. One unique aspect here is the scope of this model; while most past studies which employed these techniques for numerical weather prediction have focused on a small domain, or just a sampling of points, the models trained here demonstrate an ability to generate skillful, reliable forecasts year-round for all of CONUS and a range of lead times. There are many forecast problems that remain to be explored, but the results of this study and others strongly suggest that further development and application of these data-intensive statistical techniques could substantially improve our forecasts over the current state of the art, even compared with using more sophisticated dynamical models. To that end, implementation of this methodology for operational use to assist Weather Prediction Center forecasters with the development of their excessive rainfall outlooks is currently underway. This forecast technique presents some advantages over purely dynamical approaches, as dynamical models are inherently limited by two factors by which these statistical techniques are not. First, dynamical models require ever increasing computational resources for increasing model resolution; constraints on computing power prevent sufficient resolution to directly resolve many small-scale processes, many of which are observed in the highest impact weather phenomena. Second, dynamical models are limited by our physical understanding of the processes we are attempting to simulate or forecast. Machine learning algorithms, in contrast, can detect predictive patterns in the available information even in places where we do not know or understand the physical connection between the information and the phenomenon which we wish to predict. While they are also lim-

ited in complexity by computational and data resources, the strict limits on resolvability are not there: physical resolution can often be gained through post-processing of larger scale information. There is thus ample reason to believe that further investigation of these techniques for NWP is a worthwhile venture, and eventual implementation into forecast operations could help forecasters with their tasks by skillfully synthesizing many different sources of forecast information to help alleviate their often time-pressed schedules. This in turn can aid end-user preparedness and, in the case of high-impact events, hopefully help to protect lives and property.

One of the main advantages of the methods explored in this study compared with other popular machine learning methods, in addition to their computational tractability, is the ability to visualize their output and gain insights into detecting and quantifying specific biases in the underlying GEFS/R model, and physical insights into the most valuable forecast information for predicting locally extreme precipitation. For reasons of focus and brevity, the diagnostics that shed these insights have been omitted from this manuscript and are presented instead in a companion paper focused on the diagnostics rather than the forecasts and forecast process explored in depth here.

Some limitations of this work are worthy of note. Stage IV precipitation is used as truth for this study; though there is not a clearly better verification source available, it does have its drawbacks. It does have some spurious quality control issues, and often stuggles in areas of complex terrain due to radar beam blockage, interference, and limited gauge coverage (Herman and Schumacher 2016a; Nelson et al. 2016). Since the model is trained to forecast Stage IV QPE exceedances, this can lead to some idiosyncrasies and other anomalies associated with the nuances and persistent characteristics of the Stage IV product. One such anomaly is the persistent presence of very small areas of exceedances in some regions of complex terrain during times of favorable convective conditions. This can be removed by quality control procedures to an extent, but some artifacts do remain. This happens most prominently in the terrain of western New Mexico, and a small region

has many more instances of ARI exceedances over both the training and test periods than any of the corresponding areas. The ML-based models recognize this, and for the SW region consistently issue much higher probabilities right around this region than in the immediate surroundings. In one sense, this is correct—it is correctly predicting what it was trained to predict—but is still undesriable behavior due to a disparity between "truth" in the study and the true extreme rainfall risk. Solutions to this issue and related issues in other parts of the country must be explored in the future so that the method realizes more operational utility. Additionally, while the choice of using the ARI framework was intentional decision and provides numerous benefits, it is not an end-all for predicting heavy precipitation impacts. While ARIs have *better* correspondence with impacts than a fixed threshold, there are still regional discrepancies in which ARIs have optimal association with impacts, and the framework employed here does not account for antecedent conditions, which can be critical for assessing flash flood risk.

Additionally, the predictors for this study come from a very coarse and otherwise rather antiquated global model. The GEFS/R was used for this study because, unlike almost any other dynamical model, it has been nearly static for a very long period of record and has nearly stationary bias characteristics—an essential property for performing this kind of analysis. However, the models trained herein are not working off of the 'state of the art' of flash flood predictors. The longer range Day 2 and Day 3 lead times were chosen for this study in part because the discrepancy between GEFS/R forecast quality and 'state of the art' is smaller at these longer lead times due to less convection-allowing guidance being available, and higher-resolution models degrading in utility with increasing forecast lead time.

There are also some complications that must be considered for real-time implementation. As one example, the regional models are trained completely independently of one another, with different training data and different solutions. Consequently, they can occasionally give rather different

predictions on nearly identical inputs, resulting in undesirable probability discontinuities across region boundaries. Appropriate methods for removing probability discontinuities in space must be further explored.

Future work will seek to alleviate these limitations in a variety of ways. Exploration of using different predictands, likely combining hydrometeorological information from a variety of sources, will be made for more explicit flash flood prediction. This may involve a regionally varying predictand definition, with some ARI thresholds better corresponding to flash flood impacts in some regions compared with others. Additionally, although a large number of predictors were explored in this study, there are many additional choices for predictors that could ostensibly further improve forecast skill. While atmospheric fields are represented here in absolute terms, it may be beneficial to instead represent some fields relative to the local climatology of the forecast point in terms of standardized anomalies. This is particularly true for fields like PWAT, where standardized anomalies have often shown better correspondence with precipitation impacts across varied regions than absolute values (e.g. Junker et al. 2009; Graham and Grumm 2010; Nielsen et al. 2015). More exploration of derived fields of physical relevance to extreme precipitation processes should also be explored. Some possible examples include upslope flow to gauge forcing for ascent by the horizontal wind, column mean wind to ascertain potential for slow-moving storms, and deep-layer shear as a metric for supercell potential.

This study also focused on a rather specific time interval and took all dynamical predictors from a single, somewhat antiquated ensemble system. Future expansion both to the 12–36 hour Day 1 period and beyond the Day 3 period will be explored, including predictors from more contemporary CAM models and potentially including observations as well for the shorter lead time forecasts. Operational models also tend to undergo periodic upgrades and thus do not remain static like the ensemble system used here. The sensitivity of ML model performance to changes

in dynamical model bias characterics that result from these upgrades is a question of considerable operational relevance and an additional factor worthy of future investigation. It was also seen that the ML models suffered to varying degrees from underconfidence and, in some instances, negative bias. Methods of probability calibration of the ML model probabilities as a final post-processing step should be explored in future work, and parameter choices reconsidered in light of this additional calibration. Finally, this study only explored a subset of available machine learning algorithms. Other choices, including adaptive learning algorithms, may be able to better exploit predictor-predictand relationships, appropriately update to reflect changes in an underlying dynamical model, and produce superior forecasts for the locally extreme precipitation and flash flood forecast problem.

Acknowledgments. The authors wish to thank Josh Hacker, David John Gagne, and two anonymous reviewers for insights and feedback that greatly improved the quality of the study. The authors would also like to thank Tom Hamill and Gary Bates for generating and providing the GEFS/R data which made this research possible. Erik Nielsen provided helpful assistance in the creation of Figure 3. We also wish to thank Diana Stovern, Sarah Perfater, Benjamin Albright, Mark Klein, Michael Erickson, and James Nelson at the Weather Prediction Center which helped improve the operational utility of this research. Funding for this research was supported by NOAA Award NA16OAR4590238 and NSF Grant ACI-1450089.

APPENDIX A

Random Forests and Their Parameters

As noted in the main text, RFs are simply an ensemble of decision trees. Decision trees consist of a network of two types of nodes: decision nodes and leaf nodes. Decision nodes each have

exactly two children, which may be either decision nodes or leaf nodes, with a binary split based on the numeric value of a single input predictor determining whether to traverse to the left or right child. A leaf node has no children and instead, makes a categorical prediction of the outcome of the input example based on the leaf's relationship to its ancestor nodes. For a given forecast, one begins at a decision tree's root, traversing through its children based on the relative value of the forecast's predictors to each decision node's threshold critical value for the predictor associated with the node. This process is repeated until a leaf node is reached; its value corresponding to the leaf becomes the tree's deterministic prediction.

Decision trees can be a powerful approach for a wide array of applications, but they also have several significant drawbacks. In particular, they are very prone to overfitting (e.g. Brodley and Utgoff 1995), fitting to the noise of the training data rather than just the underlying relationships. They also don't convey any information about forecast uncertainty, as would be the case in a probabilistic framework. RFs are used instead to alleviate these concerns by producing a probabilistic forecast in a way that can significantly decrease error from overfitting the supplied training error with only a slight increase to error from oversimplistic model assumptions, provided the trees are sufficiently uncorrelated. The difficulty then revolves around generating a large set (forest) of skillful decision trees that are not strongly correlated. The decision tree generating procedure described above is deterministic: a given set of training data will always produce the same decision tree. A forest of identical decision trees, of course, adds no value over using a single decision tree. Two additional proccesses—tree bagging and feature bagging—are employed to produce unique trees. Tree bagging produces unique trees through a straightforward bootstrapping procedure. Specifically, a forest of size B is formed from the n training examples by creating B samples of size n, with replacement, from the original training data, and running the decision tree algorithm on each sample. Overfitting due to correlated trees can still occur under this approach, particularly

if a small subset of the original features are much more robust predictors of the verifying category than the rest (Breiman 2001; Murphy 2012). To overcome this problem, feature bagging is also employed, whereby only a random subset of the m original input predictors are considered at each decision node; the size of the random subset is denoted here as Z; $1 \leq Z \leq m$. This combination can result in a set of B largely uncorrelated trees, each of which is individually fairly skillful.

With any machine learning algorithm, there are numerous considerations in the actual model construction, which manifest themselves in tunable parameters. Compared with other machine learning algorithms, such as gradient boosting or support vector machines, RFs are often praised for their relative insensitivity to their parameters with respect to model performance, but it is nevertheless important to explore the parameter space in order to realize the full utility of the algorithm. The forest size B is perhaps the most obvious parameter. The general relationship between model performance and B is well known and consistent across all prediction problems; it starts quite low at very low B, initially increases rapidly with increasing B, and then slowly asymptotes to some threshold performance limit as the relationships between input features have been fully explored by the forest and the inclusion of new trees becomes redundant. Larger forest sizes require more computational expense, so the goal is to select B such that it is small enough to be computationally tractable but large enough to be near the performance limit. Another parameter noted above is Z, the number of features to consider at each node split. If this number is too small, model performance may suffer from only considering irrelevant or otherwise unpredictive features in the context of the node; if Z is too large, performance will also suffer because of underdispersive trees producing an overfit forest solution. Another frequently explored parameter is the splitting criterion evaluation function. Most commonly used are either the Gini impurity or the information gain; past studies have shown that this choice is not important for many forecast problems. In this study, information gain will be used, and can be expressed for a training set T, candidate splitting

feature x_a and candidate split value v_a as:

$$IG(T, x_a, v_a) = H(T) - H(T|x_a < v_a) \tag{A1}$$

where H(T) is the so-called entropy of a tree, defined for each of the K verifying categories, with each category i having forecast probability p_i, as:

$$H(T) = -\sum_{i=1}^{K} p_i \log_2 p_i \tag{A2}$$

The chosen splitting feature and split value are selected among those considered which maximize Equation A1 (e.g. Quinlan 1986; Murphy 2012). However, there are two other parameters that have the most substantial influence on model performance. The first, denoted S, is the minimum number of training examples required to split a node. Traditionally, RFs create a leaf only once a node is 'pure', that is, all the remaining training examples associated with that node have the same labels—event outcomes. In this way, each tree makes a categorical prediction of the predictand outcome, and probabilities are generated only in counting the proportion of trees in the forest making a particular forecast. However, this can make predictions from an individual tree very susceptible to the outcome of a particular historical case, and in some cases result in substantial overfitting. Instead, by increasing S, an RF can be allowed to make 'impure' leaves; at these nodes, an individual tree makes a probabilistic prediction based on the proportion of remaining training examples exhibiting each event class rather than continuing to split based on the remaining training data. Making S too large, however, can result in underfitting—lumping data as indistinguishable when there are in fact underlying discernible distinctions between remaining training examples with different labels. The last parameter, denoted P, is not actually an RF algorithm parameter at all. When PCA is performed, there is always a question about the number of components to retain. Though there are some heuristics (e.g. North et al. 1982), there is no definitive method to know *a priori* how many retained components P will produce the most skillful forecasts (Wilks 2011).

If P is too small, valuable forecast data is discarded and predictive performance consequently suffers. However, if it is too large, the retained PCs eventually become essentially just noise, and the RF, by fitting to these predictor values in the training data, will yield an overfit model that does not generalize to unseen data. Experiments that will not be discussed herein revealed that using information gain to determine splits and letting B=1000 produced skill near that of an infinitely large forest, and skill was insensitive to modifications of these settings, including modest increases in the forest size beyond this point. However, the Z-S-P parameter space are explored for the models trained and those results are presented in Appendix C.

One final consideration concerns the handling of rare event scenarios. For rare event problems, one necessarily has many more examples of the common event class in comparison to the rare class, leaving the rare class somewhat underrepresented in the learning problem, and model fitting that is done with respect to the rare class is often too dependent on a small number of examples. An approach that has been applied with some success in past studies (e.g. Ahijevych et al. 2016) is to sample training data disproportionately from the rarer classes, so that the number of training example associated with each event class are approximately equal. A comparison between this so called "balanced" sampling and unmodified "unbalanced" sampling is also made and the results presented in Appendix C.

APPENDIX B

Logistic Regression and Other Algorithms

Other machine learning algorithms do not extrapolate well to the high dimensionality of the forecast problem explored here. While time to train a model is not of primary concern for operational forecasting, since it is performed only once (or periodically) offline, there are nevertheless some practical considerations; models that take months or longer to train would be unlikely to be re-

alistic choices, for example. The "online" forecasting component, that is, the time required to take a new forecast, input it into a trained model and receive a forecast, is of operational concern, but all of the forecast techniques considered here can produce forecasts in a matter of minutes, and the small differences are not considered to be of practical concern. Using the random forest classification heuristic of considering the square root of the total number of features at each node split (Geurts et al. 2006), the computational complexity of training a random forest of size B from N training examples with F features (N $>$ F) may be expressed as $O(B\sqrt{F}N\log(N))$, and may be readily parallelized across trees or within trees. Some algorithms are quadratic or, in the case of support vector machines (Cortes and Vapnik 1995), even cubic in the number of training examples, and do not parallelize as readily. Others, such as logistic regression (LR), are linear in the number of training examples, but require matrix multiplication yielding a computational complexity $O(NF^2)$. PCA pre-processing acts to make learning algorithms more computationally manageable; consequently, even otherwise potentially intractable approaches, such as LR with the $\sim$10,000 native features used here, become feasible after applying the PCA pre-processing step. It furthermore serves the goal of alleviating the so-called "curse of dimensionality" and reducing overfitting.

One sensitivity experiment compares model performance as a function of the model algorithm by comparing skill of forecasts produced by RFs with those produced with LR. LR is in many senses a simpler model than an RF, since the structural form of the relationship between the predictors and the predictand is predefined before training. RFs, in contrast, make few assumptions about the relationships between the predictors and the predictand, allowing more diverse diagnoses of underlying relationships. However, this lack of assumptions can result in some degree of overfitting which fundamentally cannot be mitigated through parameter tuning. As an application of the generalized linear model, LR assumes a linear predictors-predictand relationship via the logit

function and is thus in one aspect considerably more limited than an RF. In LR, a single regression equation, or K equations for a multicategory problem with K categories, is computed to represent the probability of the outcome being category k given the set of input predictors $\mathbf{x}$. In particular, verifying probabilities are computed using the softmax function:

$$P(y = k|\mathbf{x}) = \frac{e^{\mathbf{x}^\mathbf{T}\mathbf{w}_k}}{\sum_{j=1}^{K} e^{\mathbf{x}^\mathbf{T}\mathbf{w}_k}} \tag{B1}$$

In training a LR model, the goal is to determine the optimal weights $\mathbf{w}_k$ associated with each predictor in order to yield the most accurate predictions for each event class. As with RF models, LR can be prone to overfitting if unconstrained. For RFs, one aforementioned approach is to alleviate this problem is to increase the above-termed Z parameter, which stops node splitting earlier on and makes the model less tailored to the specific training data supplied to it. Complexity in LR can be thought of as being analogously represented by large weights, or regression coefficients. In order to ensure better generalizability of the trained regression equations, it is often good practice to penalize large weights through a process known as regularization. When this is done, the computation of optimal weights can be represented as a minimization problem with two terms. For 1) a matrix Y with binary elements that are non-zero if and only if training example i has associated verifying category k and 2) a model outputting a probability matrix P for each training example and category, the multinomial loss J to be minimized can be computed as:

$$J(\mathbf{Y}, \mathbf{P}(\mathbf{w})) = \frac{1}{2}\mathbf{w}^\mathbf{T}\mathbf{w} - \frac{1}{CN}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{Y}_{i,k}log(\mathbf{P}_{i,k}) \tag{B2}$$

where C represents the extent of regularization, with smaller values indicating that large weights are penalized more than with larger values of C. Alternative approaches to regularization exist (e.g. Murphy 2012), and are explored to some degree in sensitivity experiments of Appendix C.

APPENDIX C

Results: Parameter Tuning

RF model parameters were tuned for each region and lead time separately through the 4-fold cross-validation procedure employed throughout the study. Overall, the optimal parameters were found not to vary with the two different lead times, but did vary for two of the parameters as a function of forecast region, at least to an extent; the full results appear in Table C1. For the S parameter—the number of predictors considered for each node split, the default heuristic of the square root of the total number of features was found to maximize RPSS for all regions and lead times. In all instances where both were tested, unbalanced sampling from the event classes in proportion to their true observed frequencies outperformed balanced equal sampling from each event class, in contrast to Ahijevych et al. (2016) and others; the finding appeared to be attributable to biased probabilities produced from the balanced sampling technique. For the Z parameter, the minimum number of remaining training examples in an impure parameter subspace required to perform a further node split, was generally found to be around 120. Lesser values maximized skill in the western regions, with values of 30 maximizing skill in the SW and ROCK regions, and Z=4 producing the best skill over PCST. A couple of the larger regions of the east, SE and MDWST, maximized RPSS with a value of 240, although the sensitivity between Z=120 and Z=240 was small for all regions. For P in the CTL_PCA models, skill was generally maximized with P=30, that is, retaining the 30 PCs which explain the most variance of the entire GEFS/R predictor set. For most regions, there was very limited sensitivity in the P=30–40 interval—although there was larger sensitivity outside this interval—and P=40 was found to produce slightly better skill in the NGP region. The PCST region was again the main exception, where P=60 was found to maximize cross-validation RPSS.

LR model parameters were tuned using an identical framework to ascertain the type of regularization, either based on a L1 norm which penalizes non-zero weights, or L2 norm—described in Appendix B—which penalizes large magnitude weights. L2 regularization was consistently found to produce superior results, perhaps because the number of retained PCs was already taken from the P parameter in the RF experiments, acting to nullify many potential non-zero weights of higher numbered PCs. Unlike the RF experiments, there were occasionally some large differences in the obtained optimal regularization parameter value C between lead times within the same region. Generally, models performed better with more regularized solutions, but there were some notable exceptions, with the Day 2 NGP model and Day 3 NE model obtaining optimal C parameter values on the other end of the spectrum.

References

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31 (2)**, 581–599.

Alvarez, F. M., 2014: Statistical calibration of extended-range probabilistic tornado forecasts with a reforecast dataset. Ph.D. thesis, SAINT LOUIS UNIVERSITY.

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.*, **239 (1)**, 306–337.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting*. *Wea. Forecasting*, **17 (4)**, 783–799.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47 (3)**, 805–818.

970 Baars, J. A., and C. F. Mass, 2005: Performance of National Weather Service forecasts compared
971 to operational, consensus, and weighted model output statistics. *Wea. Forecasting*, **20 (6)**, 1034–
972 1047.

973 Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney,
974 2015: improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experi-
975 ment. *Bull. Amer. Meteor. Soc.*, **96 (11)**, 1859–1866.

976 Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative pre-
977 cipitation. *Mon. Wea. Rev.*, **103 (2)**, 149–153.

978 Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2006: Precipitation-
979 frequency atlas of the United States. *NOAA Atlas 14*, **2**.

980 Bonnin, G. M., D. Todd, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2004: Precipitation-
981 frequency atlas of the United States. *NOAA Atlas 14*, **1**.

982 Bougeault, P., and Coauthors, 2010: The thorpex interactive grand global ensemble. *Bull. Amer.*
983 *Meteor. Soc.*, **91 (8)**, 1059–1072.

984 Breiman, L., 2001: Random forests. *Mach. Learn.*, **45 (1)**, 5–32, doi:10.1023/A:1010933404324.

985 Brimelow, J. C., G. W. Reuter, R. Goodson, and T. W. Krauss, 2006: Spatial forecasts of maximum
986 hail size using prognostic model soundings and HAILCAST. *Wea. Forecasting*, **21**, 206–219,
987 doi:10.1175/WAF915.1.

988 Brodley, C. E., and P. E. Utgoff, 1995: Multivariate decision trees. *Machine learning*, **19 (1)**,
989 45–77.

990 Brooks, H. E., and D. J. Stensrud, 2000: Climatology of heavy rain events in the United States
991 from hourly precipitation observations. *Mon. Wea. Rev.*, **128 (4)**, 1194–1201.

992 Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of
993 precipitation using the ecmwf ensemble prediction system. *Wea. Forecasting*, **14 (2)**, 168–189.

994 Calianno, M., I. Ruin, and J. J. Gourley, 2013: Supplementing flash flood reports with impact
995 classifications. *J. Hydrol.*, **477**, 1–16.

996 Clark, A. J., W. A. Gallus Jr, and T.-C. Chen, 2007: Comparison of the diurnal precipitation
997 cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*,
998 **135 (10)**, 3456–3473.

999 Clark, A. J., W. A. Gallus Jr, and M. L. Weisman, 2010: Neighborhood-based verification of
1000 precipitation forecasts from convection-allowing NCAR WRF model simulations and the oper-
1001 ational NAM. *Wea. Forecasting*, **25 (5)**, 1495–1509.

1002 Cortes, C., and V. Vapnik, 1995: Support-vector networks. *Mach. Learn.*, **20 (3)**, 273–297, doi:
1003 10.1007/BF00994018.

1004 Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier, and J. D. Tuttle, 2003: Coherence
1005 of warm-season continental rainfall in numerical weather prediction models. *Mon. Wea. Rev.*,
1006 **131 (11)**, 2667–2679.

1007 Delrieu, G., and Coauthors, 2005: The catastrophic flash-flood event of 8–9 September 2002 in
1008 the Gard Region, France: a first case study for the cévennes–vivarais mediterranean hydrome-
1009 teorological observatory. *J. Hydrometeor.*, **6 (1)**, 34–52.

1010 Duda, J. D., and W. A. Gallus, 2013: The impact of large-scale forcing on skill of simulated
1011 convective initiation and upscale evolution with convection-allowing grid spacings in the WRF*.
1012 *Wea. Forecasting*, **28 (4)**, 994–1018.

Eckel, F. A., 2003: Effective mesoscale, short-range ensemble forecasting. Ph.D. thesis, University of Washington.

Friedman, J. H., 1997: On bias, variance, 0/1loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, **1 (1)**, 55–77.

Fritsch, J. M., and R. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85 (7)**, 955–965, doi:10.1175/BAMS-85-7-955.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29 (4)**, 1024–1043.

Gagne, D. J., II, A. McGovern, J. Brotzge, M. Coniglio, J. Correia Jr, and M. Xue, 2015: Day-ahead hail prediction integrating machine learning with storm-scale numerical weather models. *AAAI*, 3954–3960.

Geurts, P., D. Ernst, and L. Wehenkel, 2006: Extremely randomized trees. *Machine learning*, **63 (1)**, 3–42.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11 (8)**, 1203–1211, doi:10.1175/1520-0450(1972)011⟨1203:TUOMOS⟩2.0.CO;2.

Gochis, D., and Coauthors, 2015: The great Colorado flood of September 2013. *Bull. Amer. Meteor. Soc.*, **96 (9)**, 1461–1487.

1035 Gourley, J. J., J. M. Erlingis, Y. Hong, and E. B. Wells, 2012: Evaluation of tools used for moni-
1036 toring and forecasting flash floods in the United States. *Wea. Forecasting*, **27 (1)**, 158–173.

1037 Graham, R. A., and R. H. Grumm, 2010: Utilizing normalized anomalies to assess synoptic-scale
1038 weather events in the western united states. *Wea. Forecasting*, **25 (2)**, 428–445.

1039 Grams, J. S., W. A. Gallus Jr, S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The
1040 use of a modified Ebert-McBride technique to evaluate mesoscale model QPF as a function of
1041 convective system morphology during IHOP 2002. *Wea. Forecasting*, **21 (3)**, 288–306.

1042 Hall, T., H. E. Brooks, and C. A. Doswell III, 1999: Precipitation forecasting using a neural
1043 network. *Wea. Forecasting*, **14 (3)**, 338–345.

1044 Hamill, T. M., 2017: Changes in the systematic errors of global reforecasts due to an evolving data
1045 assimilation system. *Mon. Wea. Rev.*, **(2017)**.

1046 Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr, Y. Zhu,
1047 and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast
1048 dataset. *Bull. Amer. Meteor. Soc.*, **94 (10)**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

1049 Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying
1050 climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2924.

1051 Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation fore-
1052 casts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*,
1053 **143 (8)**, 3300–3309.

1054 Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on
1055 reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134 (11)**, 3209–3229.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132 (6)**, 1434–1447.

Hapuarachchi, H., Q. Wang, and T. Pagano, 2011: A review of advances in flash flood forecasting. *Hydrological Processes*, **25 (18)**, 2771–2784.

Herman, G. R., and R. S. Schumacher, 2016a: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, doi:10.1175/WAF-D-16-0093.1.

Herman, G. R., and R. S. Schumacher, 2016b: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea. Forecasting*, **31 (2)**, 467–482.

Hershfield, D. M., 1961: Technical paper no. 40: Rainfall frequency atlas of the United States. *Weather Bureau, Department of Commerce, Washington, DC*.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Mon. Wea. Rev.*, **141 (12)**, 4564–4575.

Hong, T., P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, 2016: Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, **32 (3)**, 896–913.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage iv toward cpc gauge-based analysis. *J. Hydrometeor.*, **15 (6)**, 2542–2557.

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, **5 (1)**, 128–138.

Junker, N. W., M. J. Brennan, F. Pereira, M. J. Bodner, and R. H. Grumm, 2009: Assessing the potential for rare precipitation events with standardized anomalies and ensemble guidance at the hydrometeorological prediction center. *Bull. Amer. Meteor. Soc.*, **90 (4)**, 445–453.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, doi:10.1175/WAF906.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23 (5)**, 931–952.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16 (6)**, 672–682.

Lackmann, G. M., 2013: The south-central US flood of May 2010: Present and future. **26 (13)**, 4688–4709.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136 (9)**, 3408–3424.

Lin, Y., and K. E. Mitchell, 2005: 1.2 the NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. Hydrology, American Meteorological Society, San Diego, CA, USA*, Citeseer.

Marjerison, R. D., M. T. Walter, P. J. Sullivan, and S. J. Colucci, 2016: Does population affect the location of flash flood reports? *J. Appl. Meteor. Climatol.*, **55 (9)**, 1953–1963.

Miller, J., R. Frederick, and R. Tracey, 1973: NOAA Atlas 2. *Precipitation-frequency atlas of the western United States*, **3**.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ecmwf ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122 (529)**, 73–119.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12 (4)**, 595–600.

Murphy, K. P., 2012: *Machine learning: a probabilistic perspective*. MIT press.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31 (2)**, 371–394.

Nielsen, E. R., G. R. Herman, R. C. Tournay, J. M. Peters, and R. S. Schumacher, 2015: Double impact: When both tornadoes and flash floods threaten the same place at the same time. *Wea. Forecasting*, **30**, 1673–1693, doi:10.1175/WAF-D-15-0084.1.

Nielsen, E. R., and R. S. Schumacher, 2016: Using convection-allowing ensembles to understand the predictability of an extreme rainfall event. *Mon. Wea. Rev.*, **144**, 3651–3676.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110 (7)**, 699–706.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29 (3)**, 489–504.

Ntelekos, A. A., K. P. Georgakakos, and W. F. Krajewski, 2006: On the uncertainties of flash flood guidance: Toward probabilistic forecasting of flash floods. *J. Hydrometeor.*, **7 (5)**, 896–915.

NWS, 2017a: Service change notice 17-100. National Centers for Environmental Prediction, Weather Prediction Center, [Available online at http://www.nws.noaa.gov/os/notification/scn17-100wpc_excessive_rainfall.htm].

NWS, 2017b: Summary of natural hazard statistics in the United States. National Weather Service, Office of Climate, Weather, and Water Services, [Available online at http://www.nws.noaa.gov/om/hazstats.shtml].

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res*, **12**, 2825–2830.

Perica, S., and Coauthors, 2011: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **6**.

Perica, S., and Coauthors, 2013: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **9**.

Pielke, R. A., M. W. Downton, and J. B. Miller, 2002: *Flood damage in the United States, 1926-2000: a reanalysis of National Weather Service estimates*. University Corporation for Atmospheric Research Boulder, CO.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh models ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30 (4)**, 892–913.

Quinlan, J. R., 1986: Induction of decision trees. *Machine learning*, **1 (1)**, 81–106.

Ramshaw, J. D., 1985: Conservative rezoning algorithm for generalized two-dimensional meshes. *Journal of Computational Physics*, **59 (2)**, 193–199.

1142 Reed, S., J. Schaake, and Z. Zhang, 2007: A distributed hydrologic model and threshold
1143 frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrol-*
1144 *ogy*, **337 (3)**, 402–420.

1145 Ross, D. A., J. Lim, R.-S. Lin, and M.-H. Yang, 2008: Incremental learning for robust visual
1146 tracking. *International Journal of Computer Vision*, **77 (1-3)**, 125–141.

1147 Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2014: Climatological characteristics of atmospheric
1148 rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142 (2)**,
1149 905–921.

1150 Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation fore-
1151 casts by fitting censored, shifted gamma distributions*. *Mon. Wea. Rev.*, **143 (11)**, 4578–4596.

1152 Schmidt, J. A., A. Anderson, and J. Paul, 2007: Spatially-variable, physically-derived flash flood
1153 guidance. *AMS 21st Conference on Hydrology, San Antonio, TX B*, Vol. 6.

1154 Schumacher, R. S., A. J. Clark, M. Xue, and F. Kong, 2013: Factors influencing the develop-
1155 ment and maintenance of nocturnal heavy-rain-producing convective systems in a storm-scale
1156 ensemble. *Mon. Wea. Rev.*, **141 (8)**, 2778–2801.

1157 Schumacher, R. S., and R. H. Johnson, 2006: Characteristics of US extreme rain events during
1158 1999-2003. *Wea. Forecasting*, **21 (1)**, 69–85.

1159 Schumacher, R. S., and R. H. Johnson, 2008: Mesoscale processes contributing to extreme rainfall
1160 in a midlatitude warm-season flash flood. *Mon. Wea. Rev.*, **136 (10)**, 3964–3986.

1161 Shlens, J., 2014: A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

1162 Stevenson, S. N., and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the
1163 central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea.*
1164 *Rev.*, **142 (9)**, 3147–3162.

1165 Villarini, G., W. F. Krajewski, A. A. Ntelekos, K. P. Georgakakos, and J. A. Smith, 2010: Towards
1166 probabilistic forecasting of flash floods: The combined effects of uncertainty in radar-rainfall
1167 and flash flood guidance. *J. Hydrol.*, **394 (1)**, 275–284.

1168 Vislocky, R. L., and J. M. Fritsch, 1997: Performance of an advanced MOS system in the 1996-97
1169 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.*, **78 (12)**, 2851.

1170 Wang, S.-Y., T.-C. Chen, and S. E. Taylor, 2009: Evaluations of NAM forecasts on midtropo-
1171 spheric perturbation-induced convective storms over the US northern plains. *Wea. Forecasting*,
1172 **24 (5)**, 1309–1333.

1173 Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble trans-
1174 form (ET) technique in the NCEP global operational forecast system. *Tellus A*, **60 (1)**, 62–79.

1175 Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with
1176 0-36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23 (3)**, 407–
1177 437.

1178 Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action
1179 and collaboration. *Bull. Amer. Meteor. Soc.*, **88 (4)**, 503–511.

1180 Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic press, 676 pp.

1181 Wilson, J. W., and R. D. Roberts, 2006: Summary of convective storm initiation and evolution
1182 during IHOP: Observational and modeling perspective. *Mon. Wea. Rev.*, **134 (1)**, 23–47.

Wolter, K., J. K. Eischeid, L. Cheng, and M. Hoerling, 2016: What history tells us about 2015 us daily rainfall extremes. *Bull. Amer. Meteor. Soc.*, **97 (12)**, S9–S13.

Zeng, J., and W. Qiao, 2011: Support vector machine-based short-term wind power forecasting. *Power Systems Conference and Exposition (PSCE), 2011 IEEE/PES*, IEEE, 1–8.

Zhang, F., N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64 (10)**, 3579–3594.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185, doi:10.1175/1520-0469(2003)060⟨1173:EOMCOM⟩2.0.CO;2.

LIST OF TABLES

Symbol	Description	Predictor Group	Grid
APCP	Precipitation accumulation in past (3) 6 hours	Core	Native Gaussian
CAPE	Surface-based convective available potential energy	Core	Native Gaussian
CIN	Suface-based convective inhibition	Core	Native Gaussian
MSLP	Mean sea level pressure	Core	Native Gaussian
PWAT	Total precipitable water	Core	Native Gaussian
Q2M	Specific humidity two meters above ground	Core	Native Gaussian
T2M	Air temperature two meters above ground	Core	Native Gaussian
U10	Zonal-component of 10-meter wind	Core	Native Gaussian
V10	Meridional-component of 10-meter wind	Core	Native Gaussian
Q300	Specific humidity at 300 hPa	Upper-Air Extra	1°x1°
Q500	Specific humidity at 500 hPa	Upper-Air Core	1°x1°
Q700	Specific humidity at 700 hPa	Upper-Air Extra	1°x1°
Q850	Specific humidity at 850 hPa	Upper-Air Core1	1°x1°
T250	Temperature at 250 hPa	Upper-Air Extra	1°x1°
T500	Temperature at 500 hPa	Upper-Air Core	1°x1°
T700	Temperature at 700 hPa	Upper-Air Extra	1°x1°
T850	Temperature at 850 hPa	Upper-Air Core	1°x1°
U250	Zonal-component of 250 hPa wind	Upper-Air Extra	1°x1°
U500	Zonal-component of 500 hPa wind	Upper-Air Core	1°x1°
U700	Zonal-component of 700 hPa wind	Upper-Air Extra	1°x1°
U850	Zonal-component of 850 hPa wind	Upper-Air Core	1°x1°
V250	Meridional-component of 250 hPa wind	Upper-Air Extra	1°x1°
V500	Meridional-component of 500 hPa wind	Upper-Air Core	1°x1°
V700	Meridional-component of 700 hPa wind	Upper-Air Extra	1°x1°
V850	Meridional-component of 850 hPa wind	Upper-Air Core	1°x1°
W850	Vertical velocity (omega) at 850 hPa	Upper-Air Core	1°x1°

TABLE 1. Summary of dynamical model fields examined in this study, including the abbreviated symbol to which each variable is referred throughout the paper, a description of each variable, the predictor group with which the field is associated in the manuscript text, and the highest resolution for which the field can be obtained from the GEFS/R.

Symbol	Description
RPT1_LOCAL_MEDIAN	Median of 1-year RPTs whose closest GEFS/R grid point is the forecast point.
RPT1_LOCAL_MIN	Minimum of 1-year RPTs whose closest GEFS/R grid point is the forecast point.
RPT1_LOCAL_MAX	Maximum of 1-year RPTs whose closest GEFS/R grid point is the forecast point.
RPT10_LOCAL_MEDIAN	Median of 10-year RPTs whose closest GEFS/R grid point is the forecast point.
RPT10_LOCAL_MIN	Minimum of 10-year RPTs whose closest GEFS/R grid point is the forecast point.
RPT10_LOCAL_MAX	Maximum of 10-year RPTs whose closest GEFS/R grid point is the forecast point.
RPT1_REGIONAL_MEDIAN	Median of 1-year RPTs that lie within the domain from which model predictors are drawn.
RPT1_REGIONAL_MIN	Minimum of 1-year RPTs that lie within the domain from which model predictors are drawn.
RPT1_REGIONAL_MAX	Maximum of 1-year RPTs that lie within the domain from which model predictors are drawn.
RPT10_REGIONAL_MEDIAN	Median of 10-year RPTs that lie within the domain from which model predictors are drawn.
RPT10_REGIONAL_MIN	Minimum of 10-year RPTs that lie within the domain from which model predictors are drawn.
RPT10_REGIONAL_MAX	Maximum of 10-year RPTs that lie within the domain from which model predictors are drawn.
LAT	Latitude of forecast point.
LON	Longitude of forecast point.

TABLE 2. List of background predictors used in this study, and their associated symbols and descriptions.

Model Name	**CTL_NPCA**	**CTL_PCA**	UAC_PCA	UAF_PCA	CORE_CNFDB	CORE_CTRL	CORE_LSPACE	CORE_LTIME	**CTL_LR**
Algorithm	RF	RF	RF	RF	RF	RF	RF	RF	LR
PCA Pre-Processed		X	X	X	X				X
Uses Core Fields	X	X	X	X	X	X	X	X	X
Uses UAC Fields			X	X					
Uses UAE Fields				X					
Ensemble Information	MEDIAN	MEDIAN	MEDIAN	MEDIAN	CNFDB	CTRL	MEDIAN	MEDIAN	MEDIAN
Horizontal Radius	4	4	4	4	4	4	0 (a), 1 (b), 2 (c), 3 (d)	4	4
Timestep	3/6	3/6	3/6	3/6	3/6	3/6	3/6	12 (a), 6 (b*)	3/6

TABLE 3. Summary of the models trained in this study, and the corresponding names designated to the models. 'X' indicates the process is performed or the information is used; a lack of one indicates the opposite. MEDIAN corresponds to the ensemble median, CTRL corresponds to the ensemble control member's fields, and CNFDB uses the median in addition to the second-from-lowest and second-from-highest member values for each field. Horizontal radius is listed in grid boxes from forecast point; timestep denotes the number of hours between GEFS/R forecast field predictors. Slashes indicate the first number applies to the Day 2 version of the model, while the latter number applies to the Day 3 version. Letters enclosed by parentheses indicate sub-versions of models, with one parameter changed to the value adjacent to the letter. Asterisks indicate a model applies only to Day 2, and not Day 3. Otherwise, models apply to all eight forecast regions and have both Day 2 and Day 3 versions. Those models with bolded names are incorporated into the weighted blend of the final model configuration.

Region	S Parameter	Z Parameter	P Parameter
ROCK	SQRT	30	30
NGP	SQRT	120	40
MDWST	SQRT	240	30
NE	SQRT	120	30
PCST	SQRT	4	60
SW	SQRT	30	30
SGP	SQRT	120	30
SE	SQRT	240	30

Table C1. Optimal RF parameters obtained in cross-validation for the Z-S-P parameter space. SQRT indicates the square root of the total number of predictors; symbols are otherwise as described in the manuscript text. Evaluated values were 1, 2, 4, 8, 16, 30, 60, 120, 240, 480 for Z, and 20, 25, 30, 40, 50, 60, 70, 80, 90, 100 for P.

Region	Regularization	C Parameter, Day 2	C Parameter, Day 3
ROCK	L2	0.0001	0.0001
NGP	L2	10000	0.0008
MDWST	L2	0.0001	0.0464
NE	L2	2.78	10000
PCST	L2	0.0001	0.0001
SW	L2	0.359	0.0001
SGP	L2	0.0008	0.0464
SE	L2	21.54	0.0008

Table C2. Optimal LR parameters obtained in cross-validation for the C parameter and regularization type for all lead times and regions. Evaluated for C were 0.0001, 0.0008, 0.0060, 0.0464, 0.359, 2.78, 21.54, 167.8, 1291, and 10000.

LIST OF FIGURES

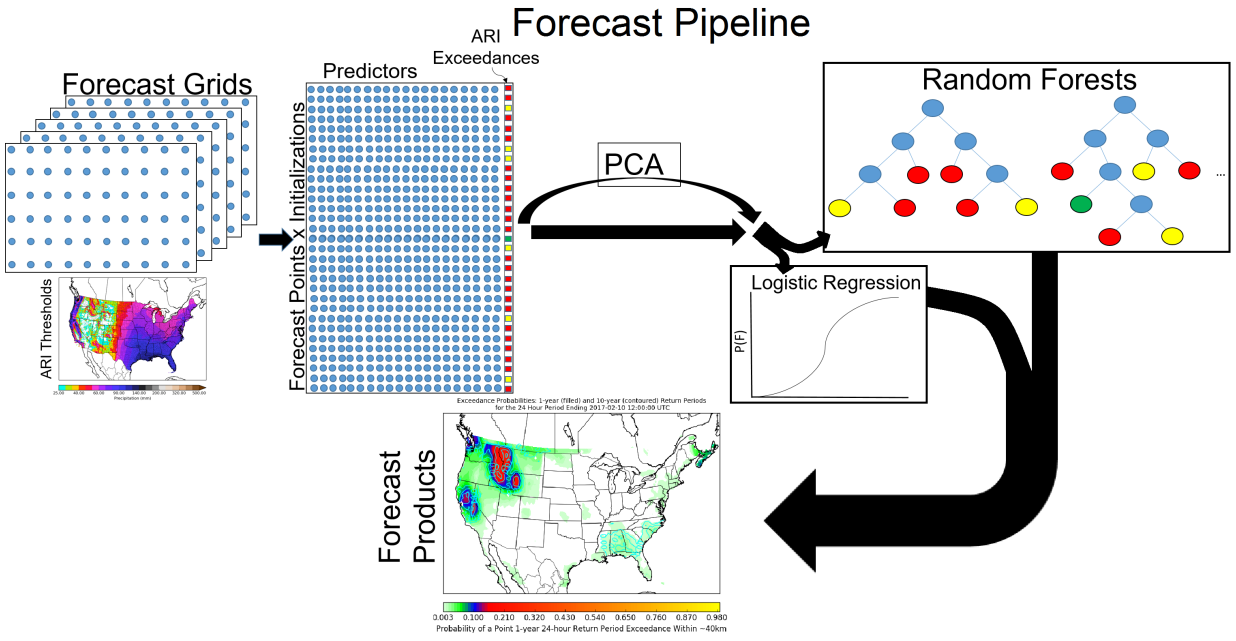
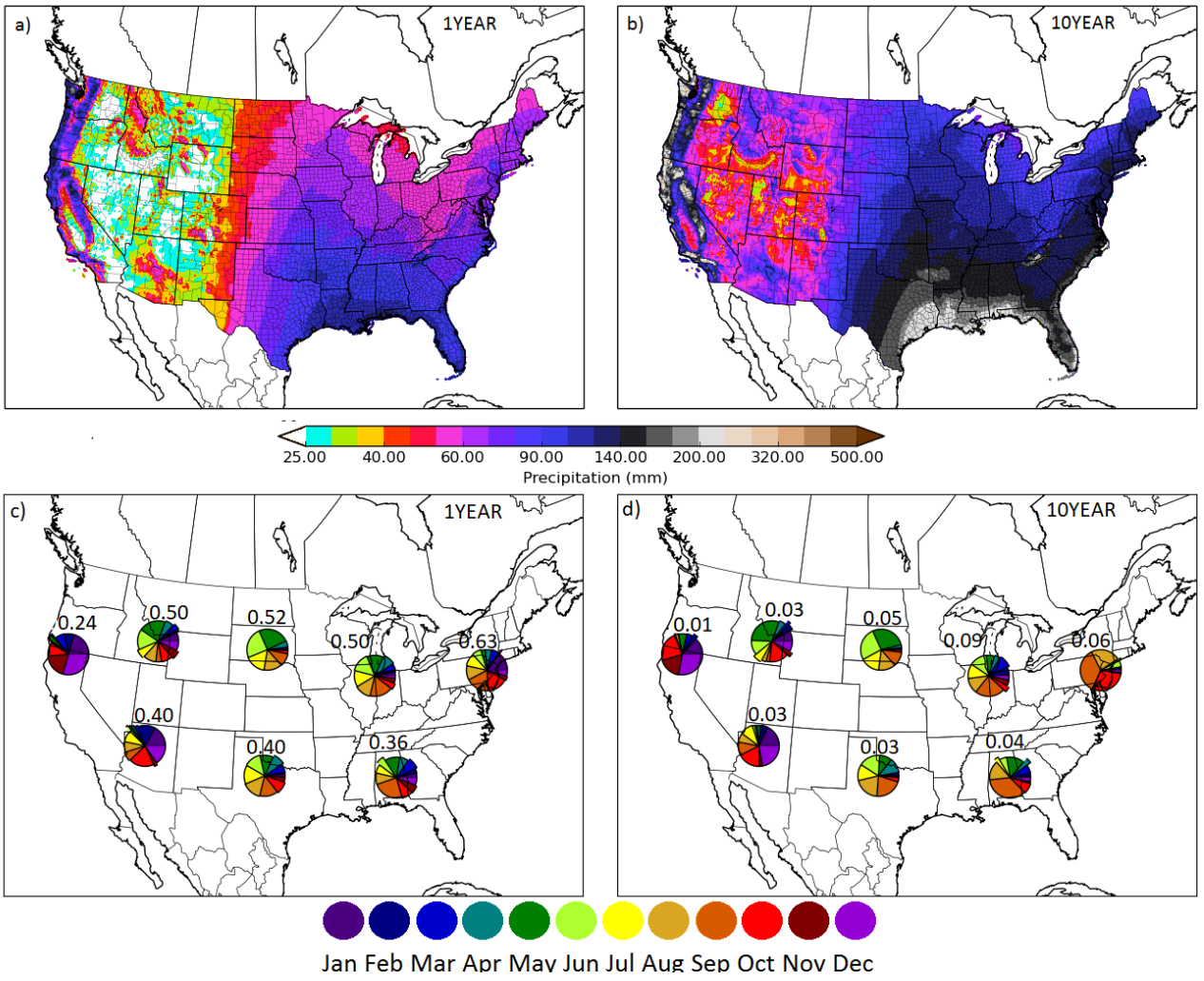
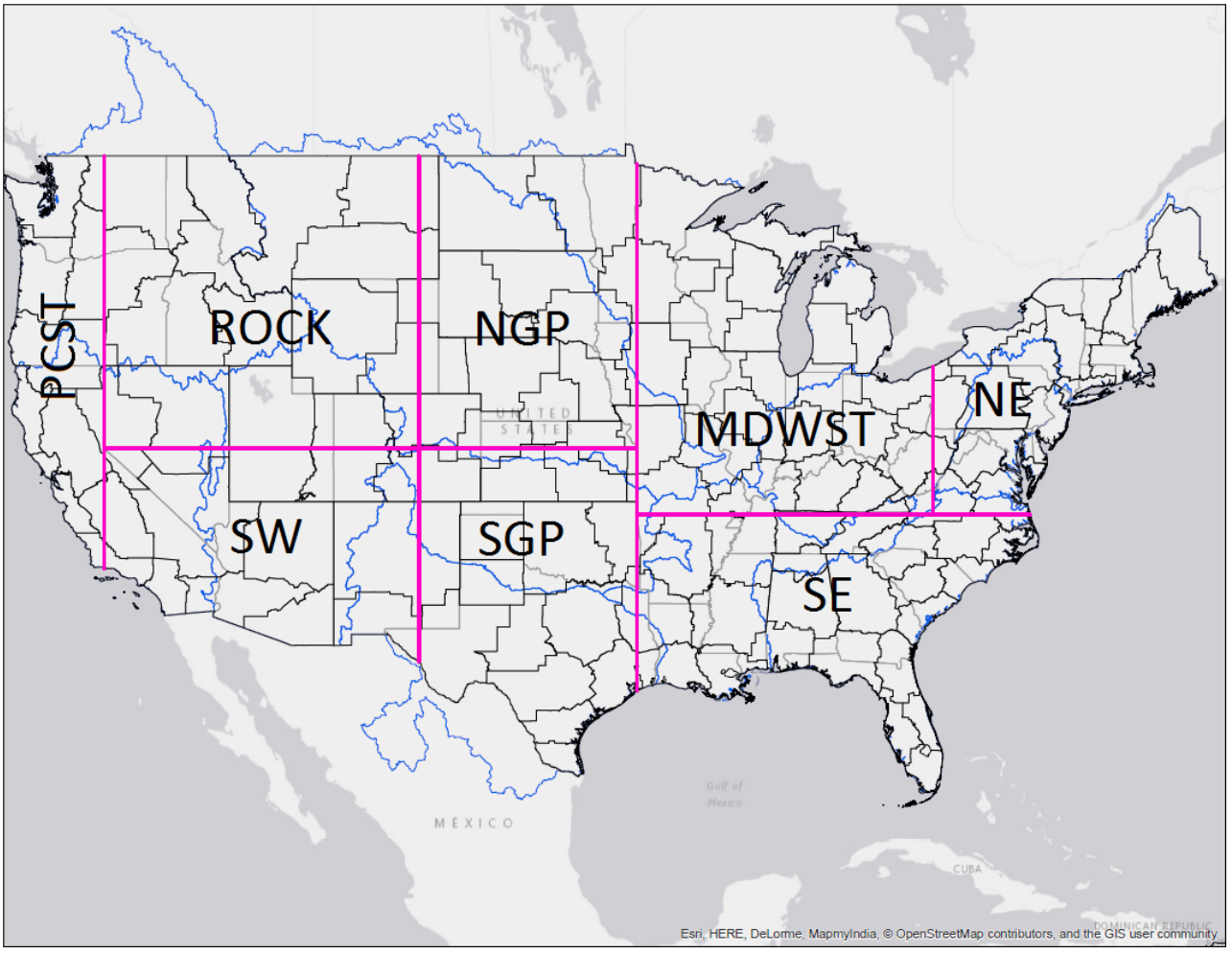


FIG. 1. Schematic representation of the forecast process for this study. GEFS/R forecasts are taken, assembled across fields, space, and time to form a training matrix, and past observations are used to associate a label with each forecast initialization, forecast day, forecast point triplet. The training matrix optionally undergoes pre-preocessing through principal component analysis, and then is input to one or more machine learning algorithms. From here, probabilistic ARI exceedance forecasts may be readily generated.



1334 FIG. 2. Return period thresholds at the (a) 1-year and (b) 10-year ARI levels over CONUS for a 24-hour
1335 accumulation interval. Climatology of observed exceedances of the (c) 1-year, 24-hour ARI thresholds and
1336 (d) 10-year, 24-hour ARI thresholds between January 2003 and August 2013 based on Stage IV Precipitation
1337 Analysis. Pie charts indicate the monthly distribution of event occurrence within each study region as shown in
1338 Figure 3. Numbers above the pie charts indicate the mean number of exceedances per point per year within the
1339 region (a priori 1 and 0.1 for 1-year and 10-year ARIs, respectively).



1340 FIG. 3. Map depicting the regional partitioning of CONUS used in this study, and the labels ascribed to each
1341 region.

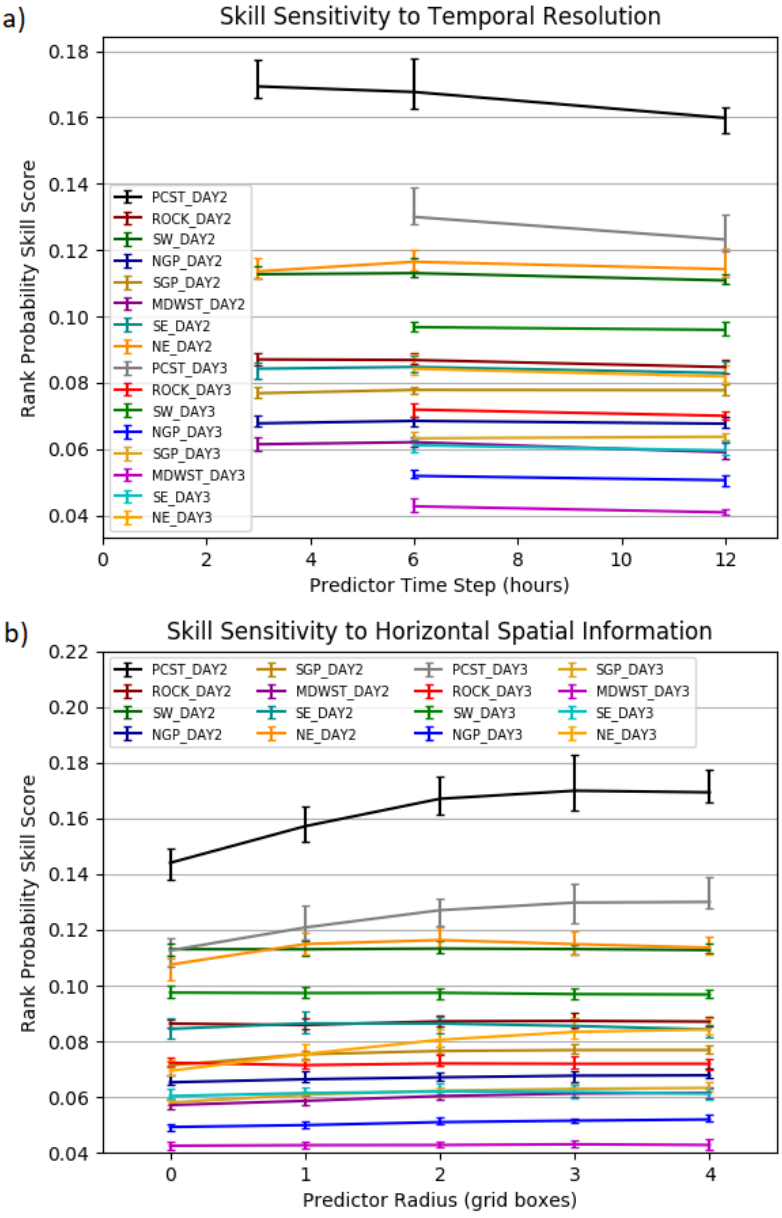


FIG. 4. Sensitivity experiment RPSS results for (a) the CORE_LTIME models, as a function of the timestep between incorporation of new atmospheric field forecast values, and (b) the CORE_LSPACE models, as a function of the radius of predictor information incorporated, both including both Day 2 and Day 3 versions of the model and for each region studied. Lines correspond to a particular day, region pair as indicated in the respective panel legends. Error bars in both panels correspond to 90% confidence bounds obtained by bootstrapping.

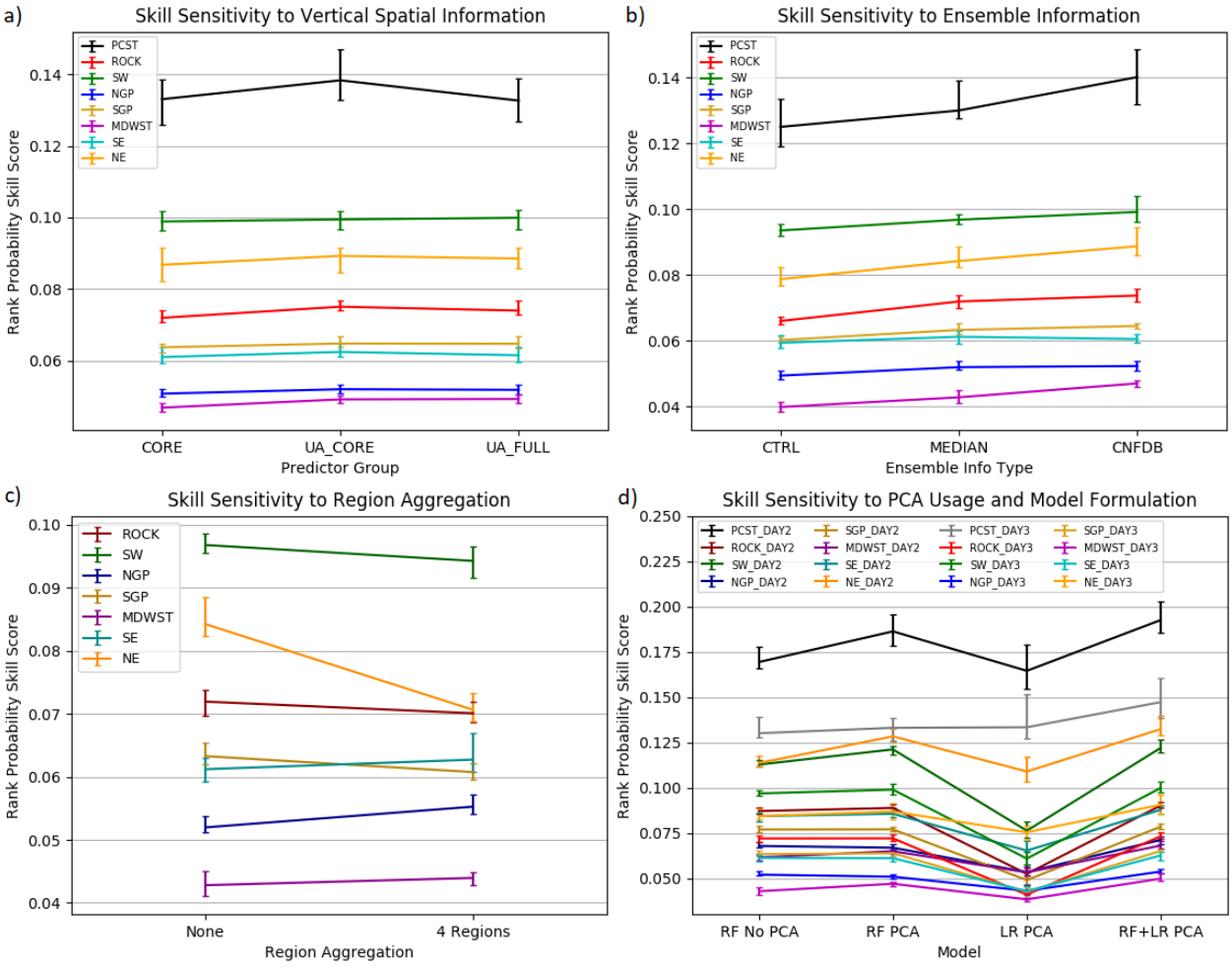


FIG. 5. Sensitivity experiment RPSS results. Panel (a) as a function of the atmospheric fields included as input to the RF algorithm, for Day 3 forecast and broken out by region. From left to right, the columns correspond to results using: 1) just the 'Core' atmospheric field group, 2) both the 'Core' and 'Upper Air Core' groups, 3) the 'Core', 'Upper Air Core', and 'Upper Air Extra' groups. For more information on which fields are included in each predictor group, consult Table 1. Panel (b) as a function of the type of GEFS/R information used as input predictors to the RF algorithm, for Day 3 forecasts and broken out by region. From left to right, the columns correspond to results using: 1) just the forecast fields from the GEFS/R control member, 2) the ensemble median forecast values from the full ensemble, 3) the ensemble median, 2nd-from-minimum, and 2nd-from-maximum forecast values from the full ensemble. Panel (c) as a function of region aggregation, with the left column using the eight regions depicted in Figure 3, and the right column using training data which aggregates data from seven of the eight original regions into three regions, as described in the text. Panel (d) as a function of model algorithm for different forecast days and regions as indicated in the figure legend. From left to right, columns correspond to results of the CTL_NPCA model, CTL_PCA model, CTL_LR model, and a weighted combination of CTL_PCA and CTL_LR models as described in the paper text. For all panels, error bars correspond to 90% confidence bounds obtained by bootstrapping.

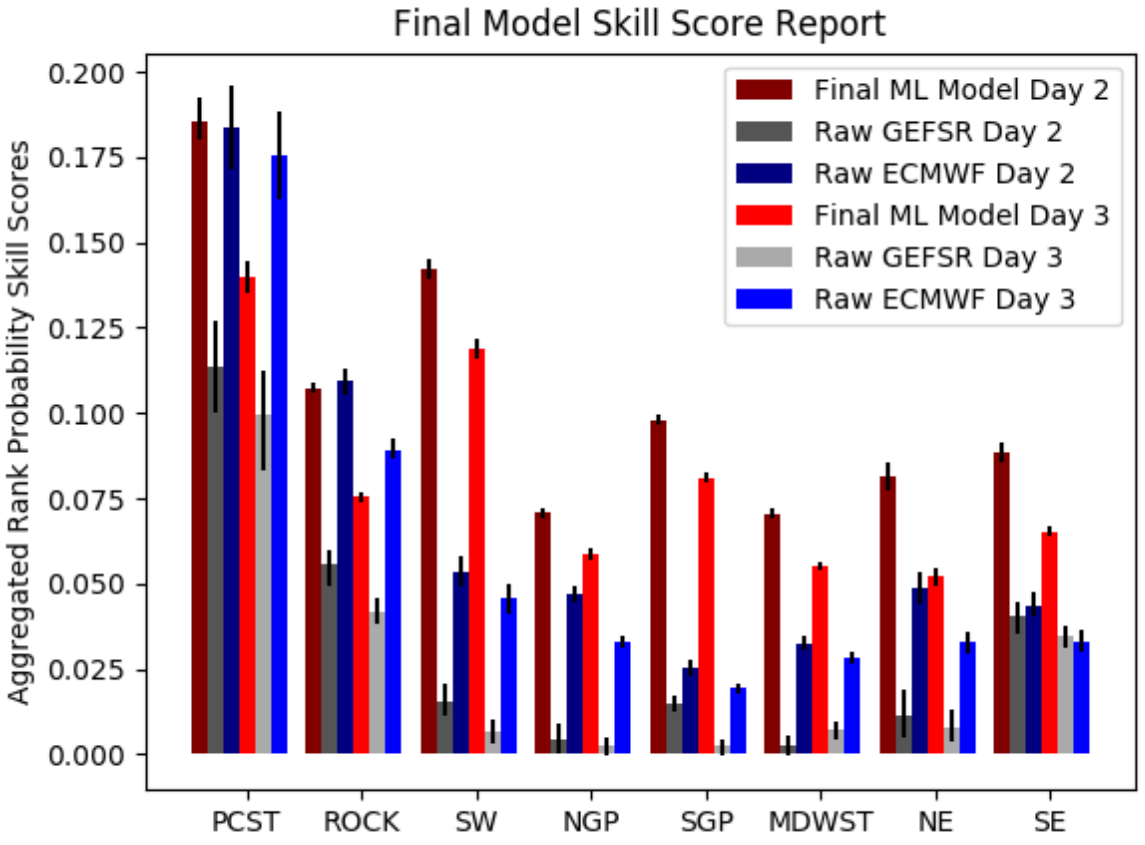


FIG. 6. Final RPSS results obtained over the four year test period spanning September 2013–August 2017, broken out by region. Red bars correspond to the results of the final forecast models trained in this study, while gray bars depict results from the raw GEFS/R QPF probabilities derived from the full ensemble. Dark bars illustrate Day 2 performance results, while lighter colors show results for Day 3. Error bars correspond to 90% confidence bounds obtained by bootstrapping.

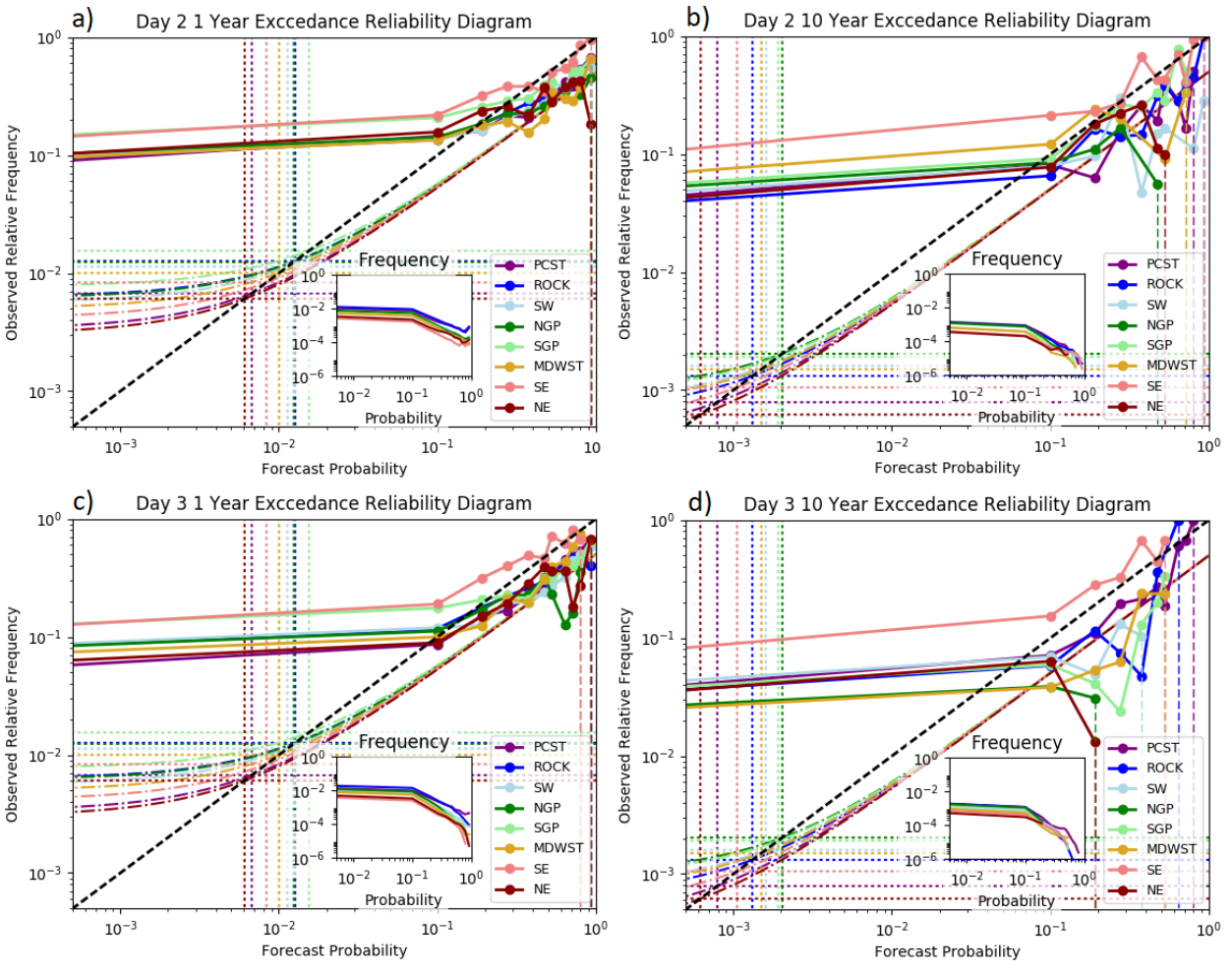


FIG. 7. Reliability diagrams for forecasts generated from raw QPFs of the full GEFS/R ensemble. Colored opaque lines with circular points indicate observed relative frequency as a function of forecast probability; the solid black line is the one-to-one line, indicating perfect reliability. Colors correspond to the performance of the forecasts over different regions, as indicated in the legend in the lower-right of each panel. Inset panels indicate the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale on the left hand side of each panel; lines are again colored by region in accordance with the legend. Panel (a) shows Day 2 forecast results for 1-year ARI exceedance forecasts, (b) to Day 2 10-year ARI exceedance forecasts, (c) to Day 3 1-year exceedance forecasts, and panel (d) to Day 3 10-year ARI exceedance forecasts. All axes are logarithmic as labeled. Colored dotted lines indicate the climatological event probability for each region for the ARI level of the corresponding panel, while the dash-dotted lines indicate no skill lines for the color-corresponding region.

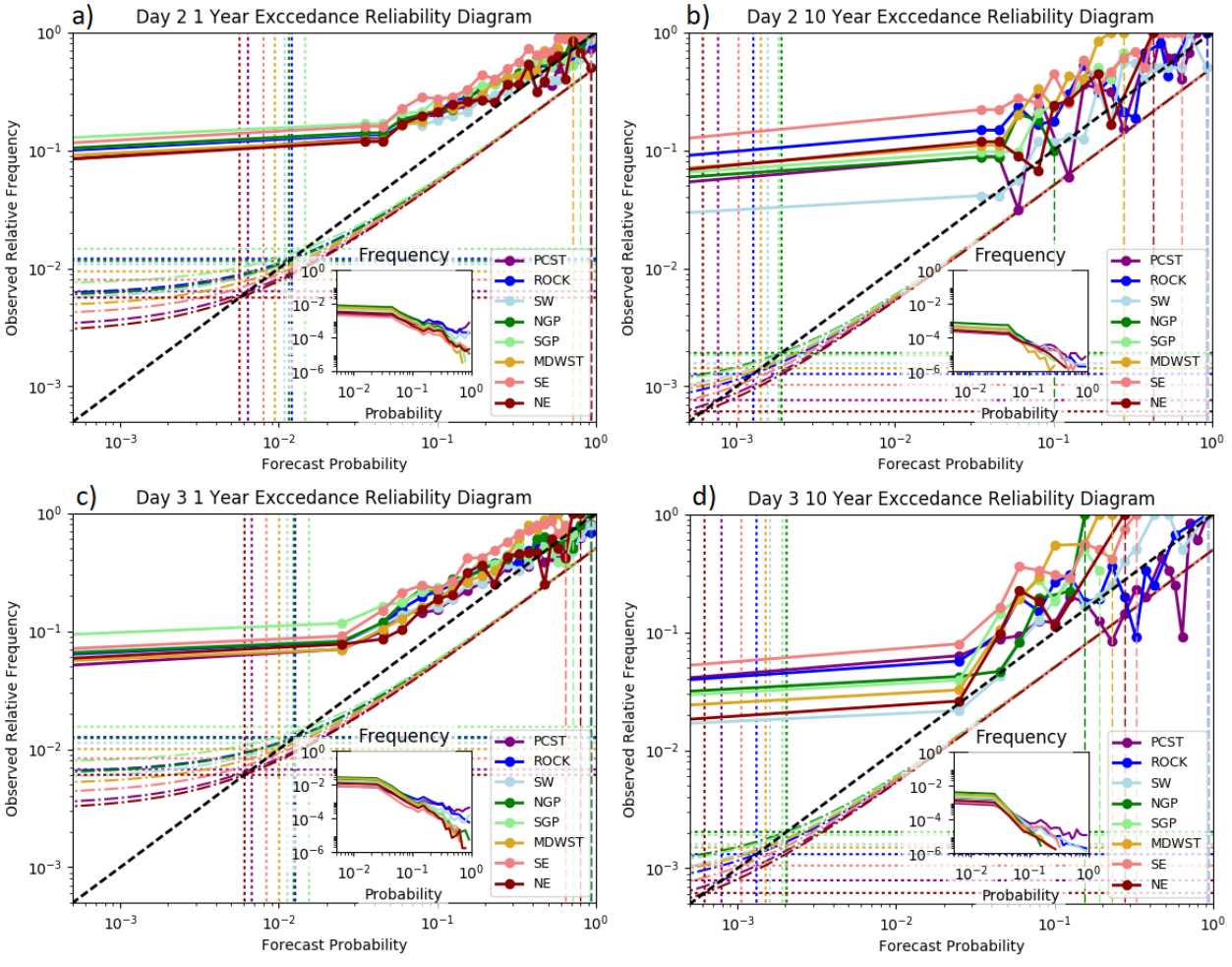


FIG. 8. Same as Figure 7, but for ECMWF ensemble forecasts.

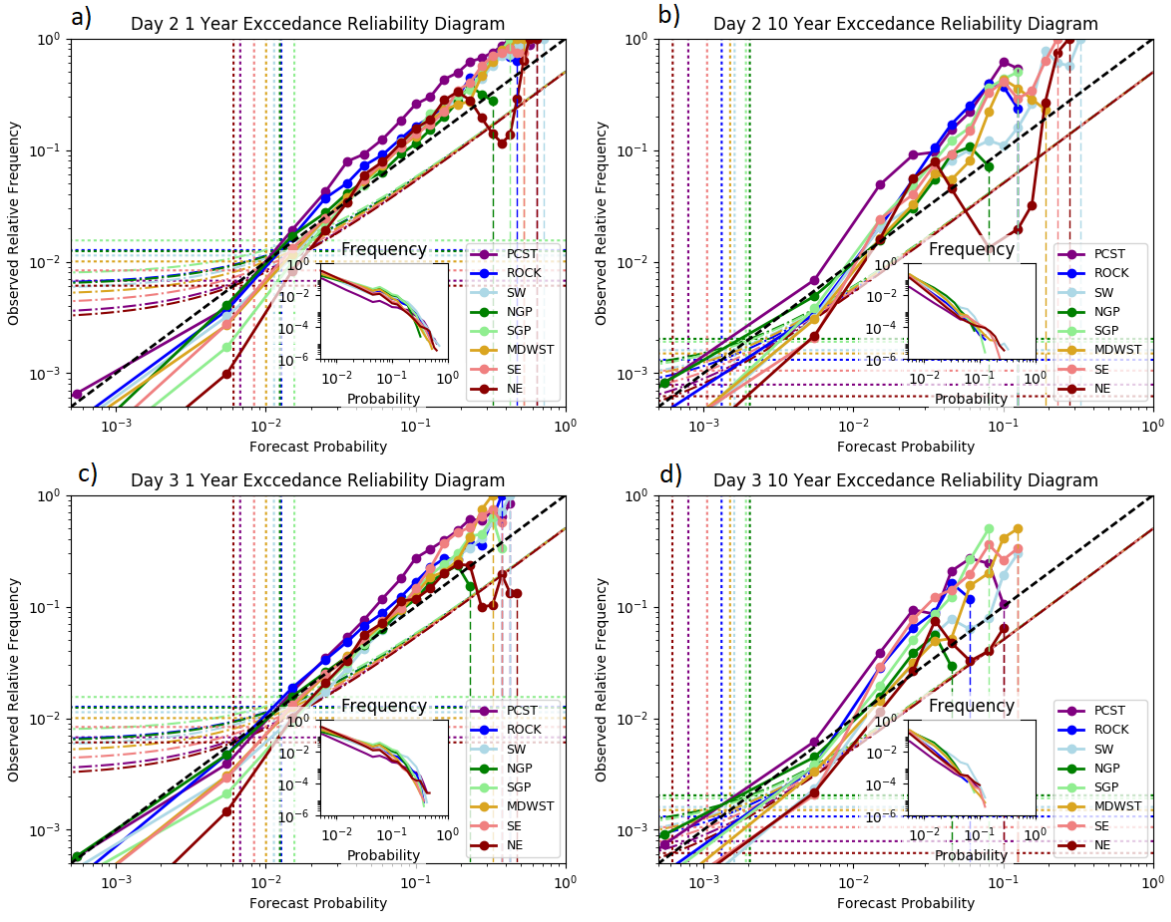


FIG. 9. Same as Figure 7, but for forecasts generated from the CTL_NPCA model.

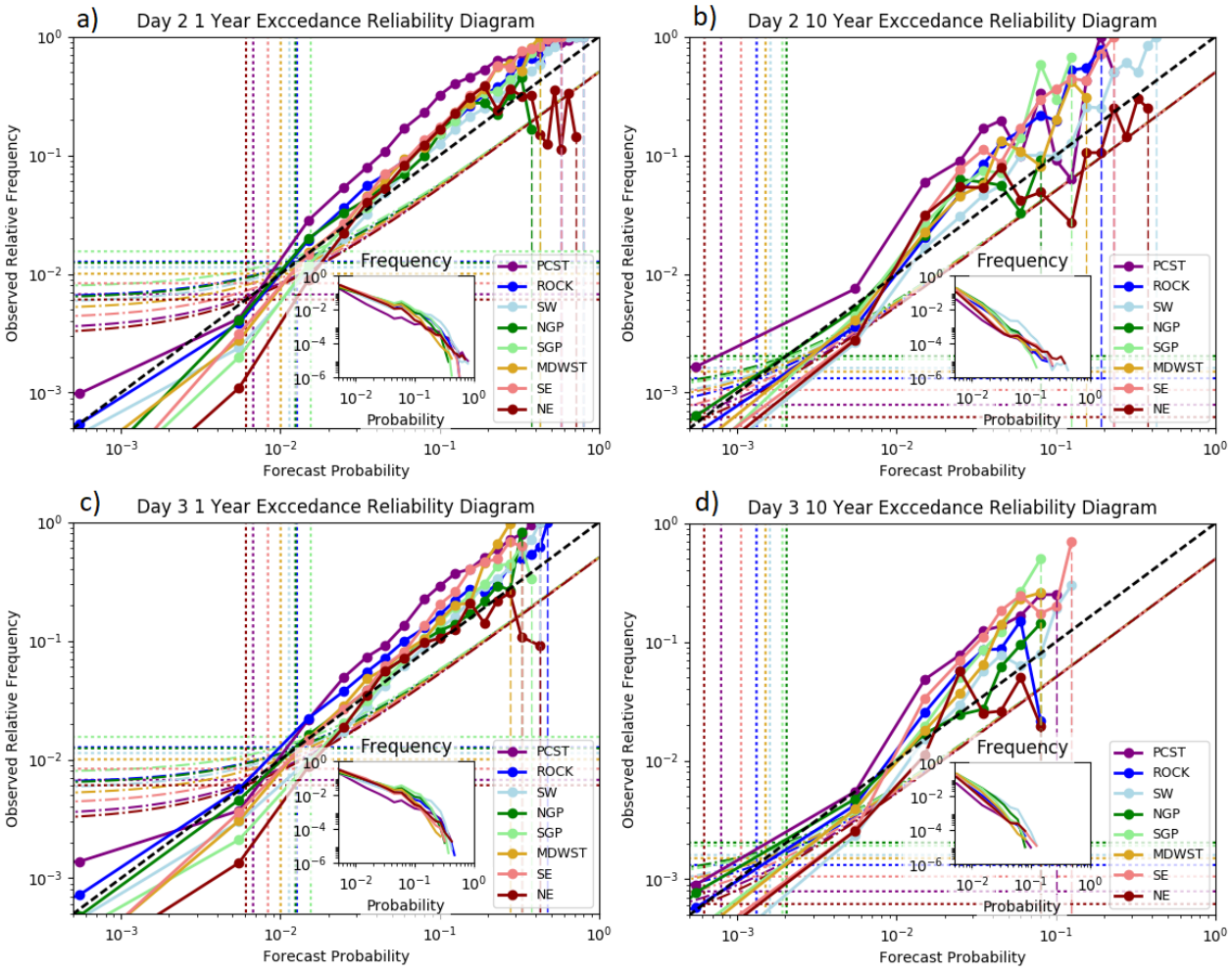


FIG. 10. Same as Figure 7, but for forecasts from the CTL_PCA model.

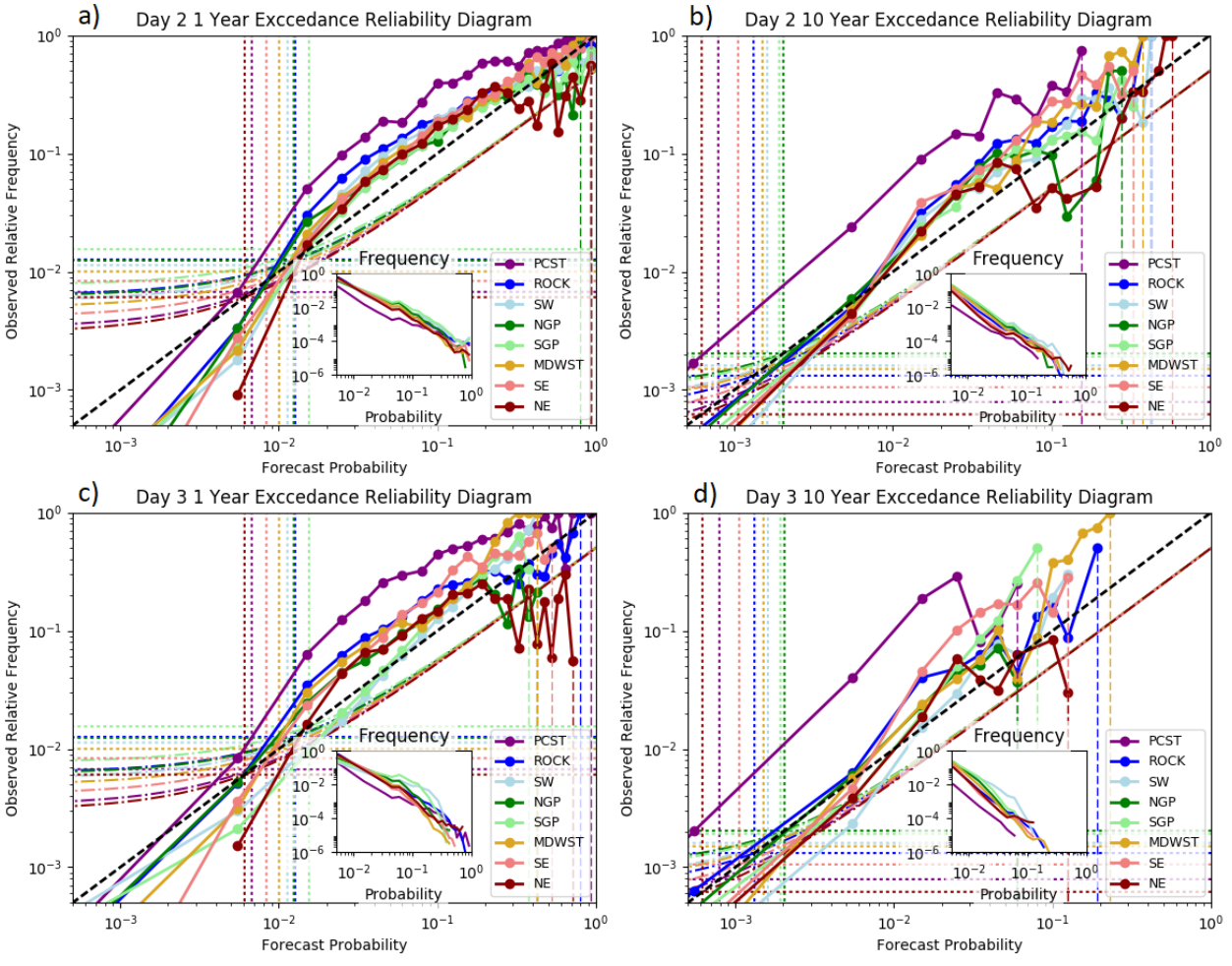


FIG. 11. Same as Figure 7, but for forecasts from the CTL_LR model.

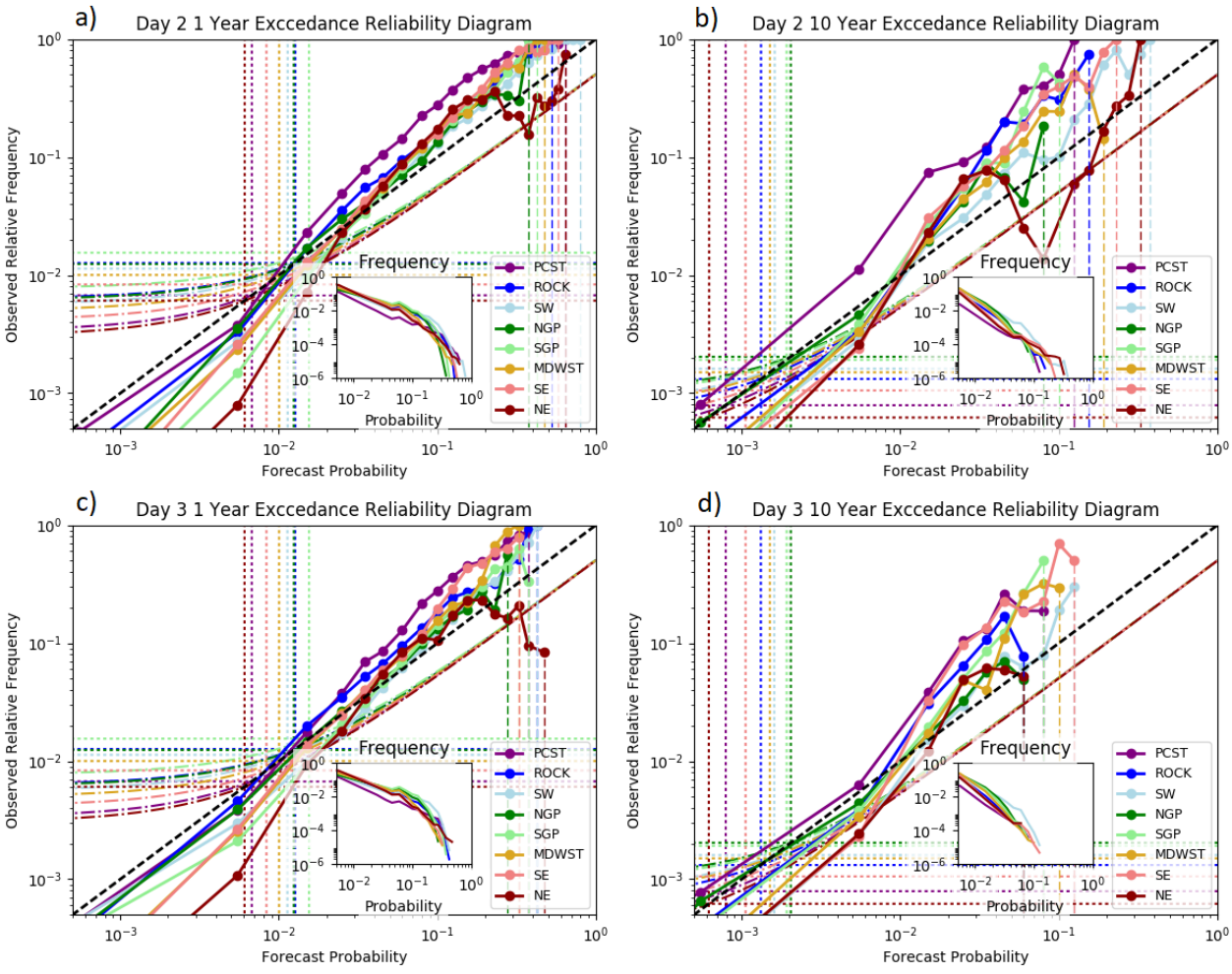


FIG. 12. Same as Figure 7, but for the final forecast model.

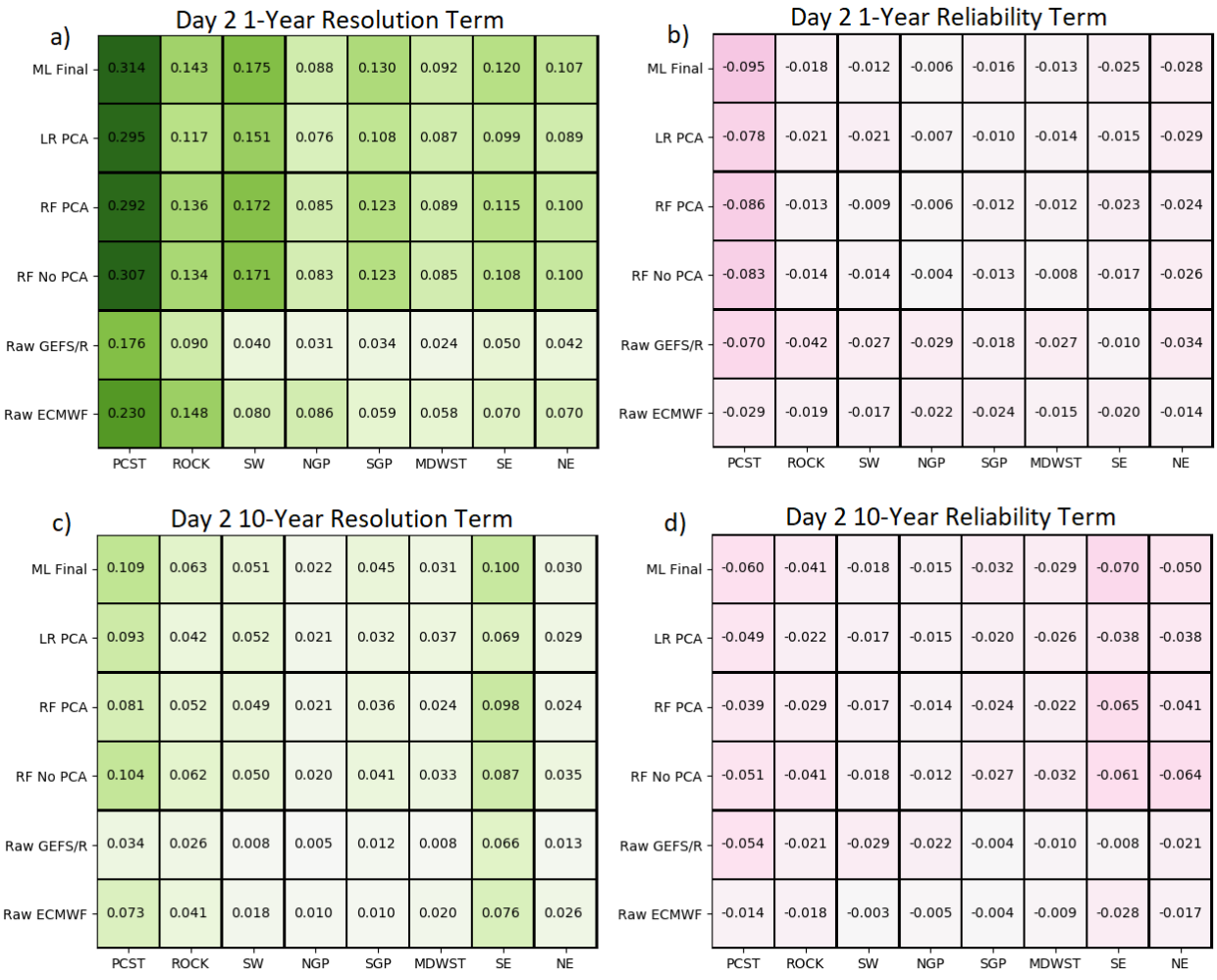


FIG. 13. Modified Murphy (1973) decomposition results, following equation 3 in text. Panel (a) depicts the equation 3 "resolution" term for all models and regions for Day 2 forecasts at the 1-year severity level, panel (b) depicts the "reliability" term results for the same forecasts and severity level. Panels (c) and (d) are analogous to panels (a) and (b), but for 10-year ARI exceedance forecasts. Numeric values indicate the value of the corresponding term of the table, as indicated by the model label (row) and region (column).

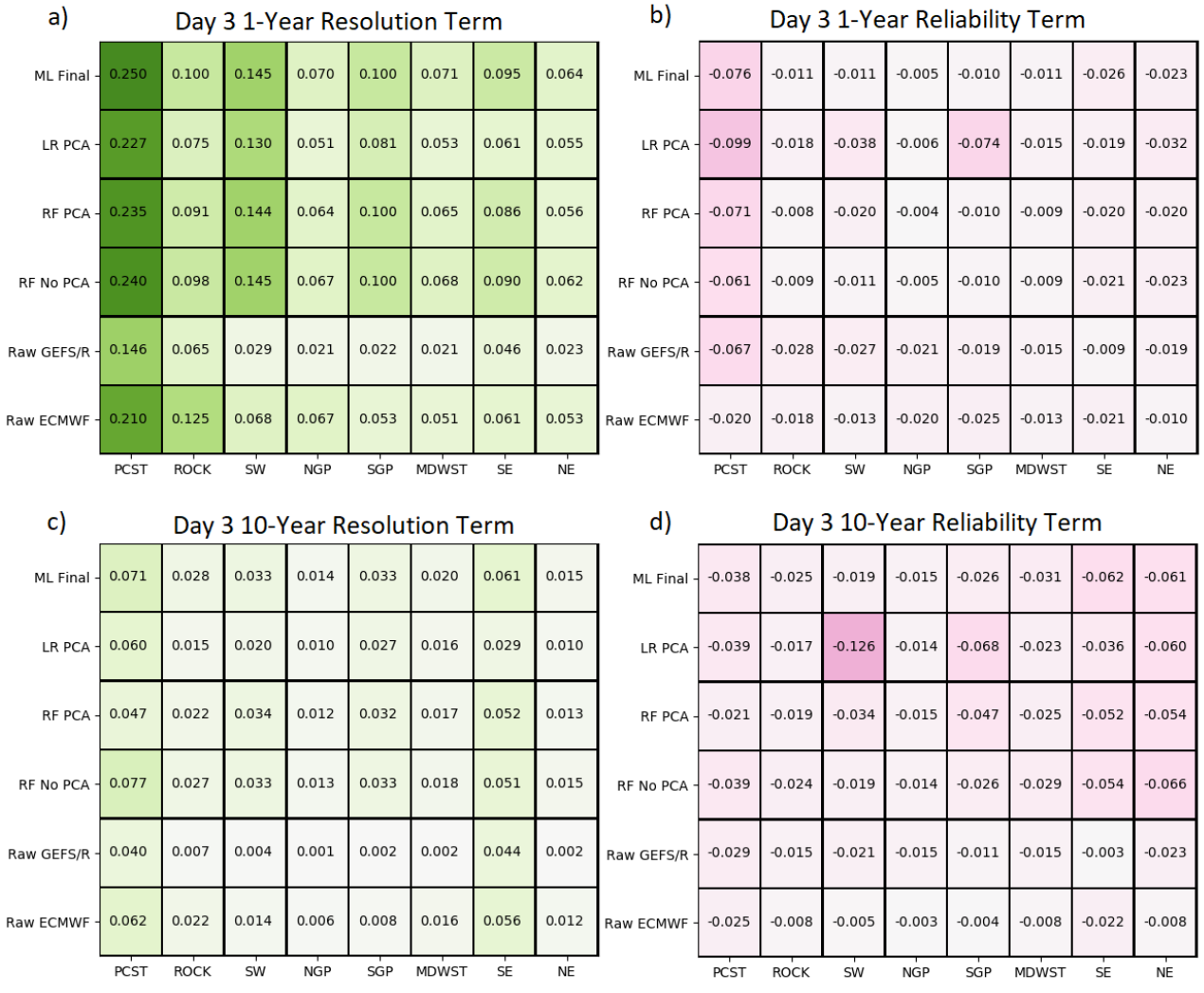


FIG. 14. Same as Figure 13, but for Day 3 forecasts.

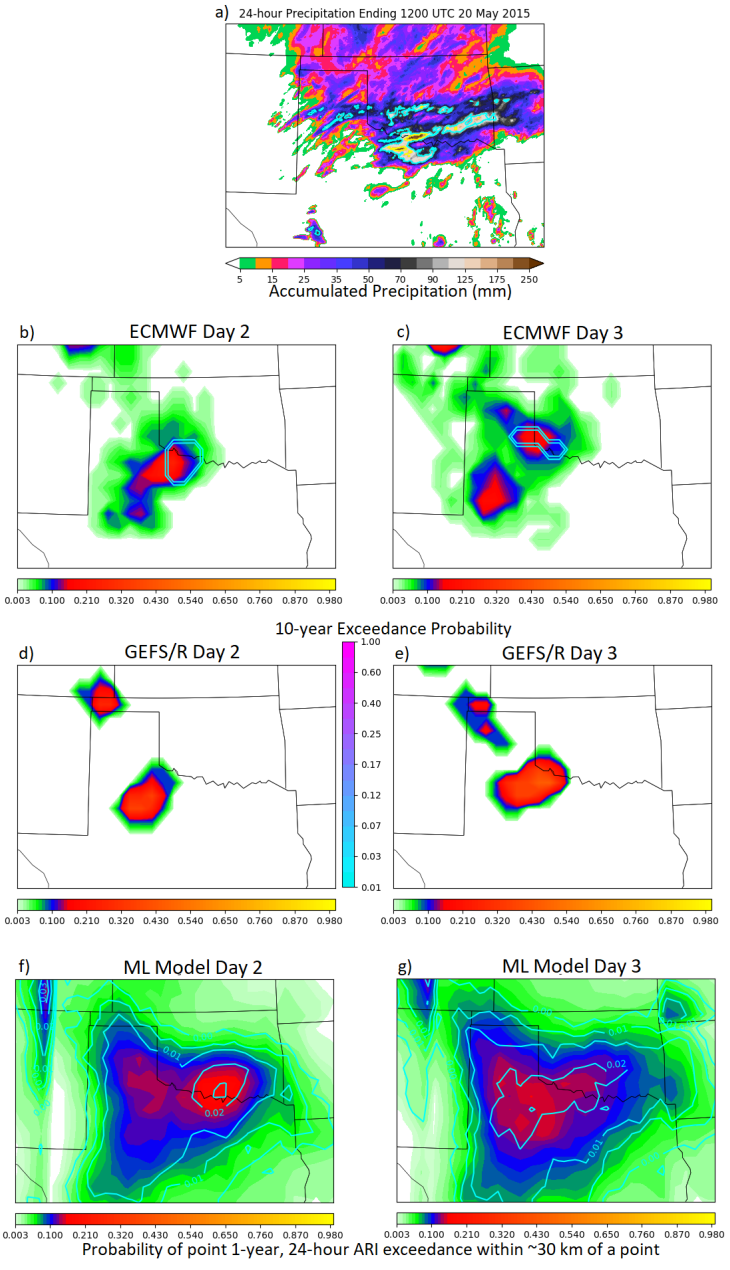


FIG. 15. Case study depicting forecasts from the final ML model and both reference ensembles for the 24-hour period ending 1200 UTC 20 May 2015. (a) 24-hour Stage IV QPE ending at 1200 UTC 20 May 2015 in filled contours. The unfilled cyan contours enclose areas with 1-year 24-hour ARI exceedances, and the unfilled yellow contours enclose areas of 10-year 24-hour ARI exceedances. (b) ECMWF ensemble neighborhood ARI exceedance probabilities in the filled (1-year) and unfilled (24-year) contours for the 36–60 hour forecast initialized 0000 UTC 18 May 2015 and (c) for the 60–84 hour forecast initialized 0000 UTC 17 May 2015. Panels (d) and (e) depict analogous fields as panels (b) and (c), respectively, except for forecasts from the raw GEFS/R QPFs. Panels (f) and (g) similarly show respectively 36–60 and 60–84 hour forecasts, except for from the final version of the ML model trained in this study.

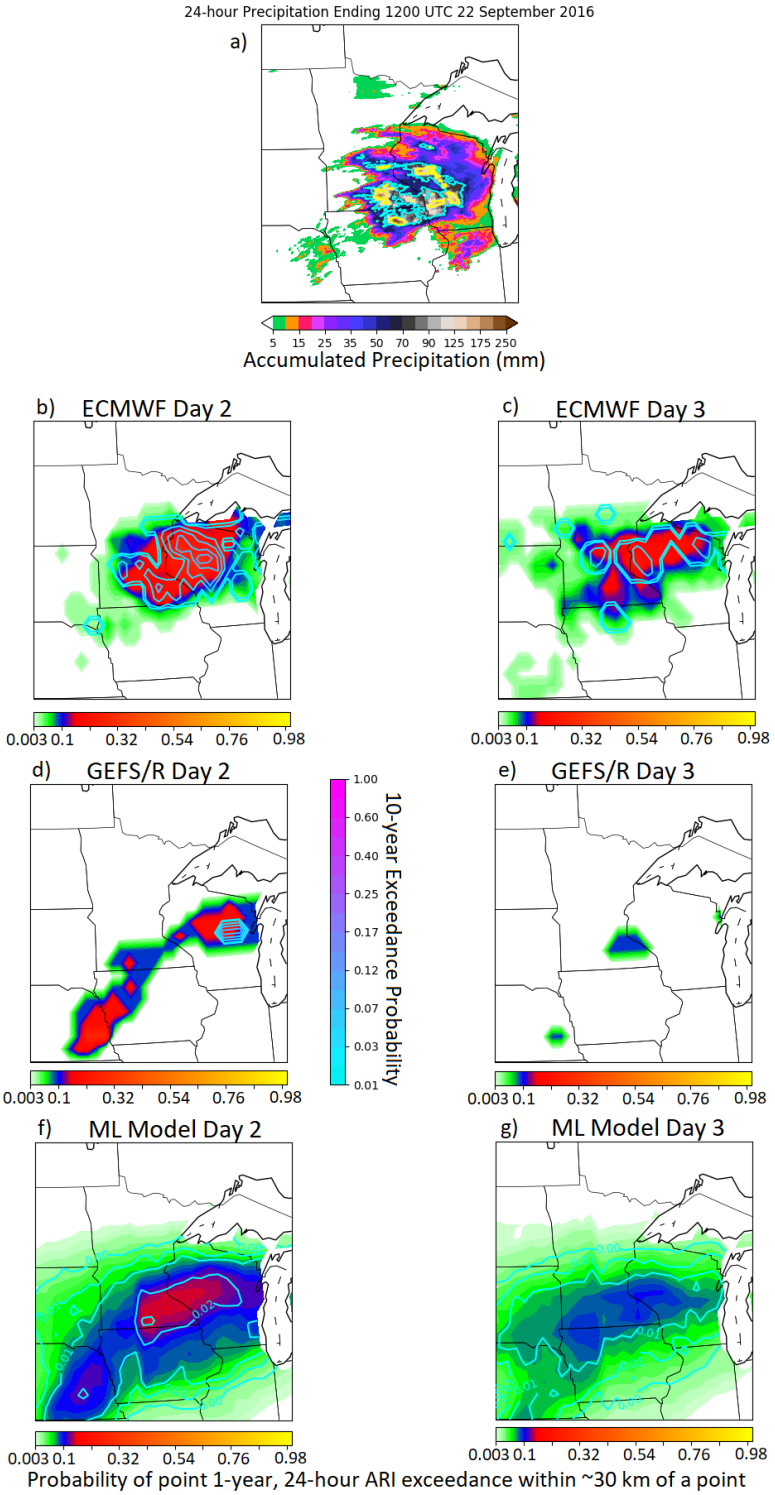


FIG. 16. Same as Figure 15, but for the 24-hour period ending 1200 UTC 22 September 2016.