"Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation

GREGORY R. HERMAN AND RUSS S. SCHUMACHER

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

(Manuscript received 24 October 2017, in final form 11 April 2018)

ABSTRACT

Three different statistical algorithms are applied to forecast locally extreme precipitation across the contiguous United States (CONUS) as quantified by 1- and 10-yr average recurrence interval (ARI) exceedances for 1200–1200 UTC forecasts spanning forecast hours 36–60 and 60–84, denoted, respectively, day 2 and day 3. Predictors come from nearly 11 years of reforecasts from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R) model and derive from a variety of thermodynamic and kinematic variables that characterize the meteorological regime in addition to the quantitative precipitation forecast (QPF) output from the ensemble. In addition to encompassing nine different atmospheric fields, predictors also vary in space and time relative to the forecast point. Distinct models are trained for eight different hydrometeorologically cohesive regions of the CONUS. One algorithm supplies the GEFS/R predictors directly to a random forest (RF) procedure to produce extreme precipitation forecasts; the second also employs RFs, but the predictors instead undergo principal component analysis (PCA), and extracted leading components are supplied to the RF. In the last algorithm, dimension-reduced predictors are supplied to a logistic regression (LR) algorithm instead of an RF. A companion paper investigated the quality of the forecasts produced by these models and other RF-based forecast models. This study is an extension of that work and explores the internals of these trained models and what physical and statistical insights they reveal about forecasting extreme precipitation from a global, convection-parameterized model.

1. Introduction

Machine learning algorithms have demonstrated considerable utility in many scientific disciplines, including computer vision (e.g., Rosten and Drummond 2006), natural language processing (e.g., Collobert et al. 2011), and bioinformatics (e.g., Larrañaga et al. 2006). Machine learning has also been used with considerable success in a wide range of future prediction scenarios, from financial market analysis (e.g., Cao and Tay 2003) to election forecasting (e.g., Bermingham and Smeaton 2011) to numerical weather prediction (NWP; e.g., Hall et al. 1999; Roebber 2013; Rozas-Larraondo et al. 2014; McGovern et al. 2017). Recently, these techniques have been receiving increasing attention and application in NWP; many of these preliminary forays have demonstrated considerable utility of these techniques over historical competitors (e.g., Herman and Schumacher 2016b), with occasional exception (e.g., Applequist et al. 2002).

One frequently noted criticism of machine learning forecast models is their lack of interpretability and neglect of underlying physics and dynamics of the forecast problem, rendering additional interpretation and analysis of their output difficult or impossible. These critiques did not first appear with the emergence of machine learning; in fact, these qualms with statistical forecast models have been expressed since the early days of NWP (e.g., Lorenz 1956). And there is legitimate reason for these concerns; given the chaotic nature of the atmosphere system, any model—statistical or dynamical—will necessarily have formulaic limitations, systematic biases, and failure modes regardless of the level of care exercised during model construction. When the model's processes are opaque, it can be difficult to rationally diagnose these circumstances, and the ability of the forecaster to add value over the raw guidance is

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-17-0307.s1.

Corresponding author: Gregory R. Herman, gherman@atmos.colostate.edu

DOI: 10.1175/MWR-D-17-0307.1

inhibited. Thus, even when, for example, a statistical model exhibits better objective performance compared with a competing dynamical model, if a human forecaster understands the underpinnings and characteristics of the dynamical model but not the statistical model, he or she may still be able to provide better final forecasts using the dynamical guidance over the statistical guidance. The ''understanding'' referenced here does not require a complete and comprehensive mathematical understanding sufficient to exactly reproduce the result by hand; even using a very simple dynamical model, it is extraordinarily difficult to reproduce an accurate forecast by manual means (e.g., Richardson 2007), and seldom are interpreters of model guidance familiar with numerical specifics, dynamical core particulars, or parameterization details. Rather, there is a well-understood overarching process of using data assimilation to produce an analysis and initialize a model that embodies the primitive equations governing the atmosphere in some capacity, then integrating the model forward in time to produce a forecast. Additionally, the intermediate steps—output from hours after initialization but before forecast valid time—are fully inspectable and comprehensible. In contrast, to many, statistical models and especially those employing machine learning seem comparatively opaque; a host of predictors are ingested and a forecast(s) is produced with little if any information provided on how the model got from the predictors it used to the answer it generated. While a small part of this is perhaps inherent to statistical forecasting, with improved visualization of statistical models developed for NWP, physical insights into how the predictors used relate to the forecasted phenomenon may be gained, and ability to deduce likely biases based on the present meteorology may be acquired.

Among statistical forecast algorithms, regression models have the longest and most extensive use in operational NWP (e.g., Glahn and Lowry 1972) and are perhaps the most easily and directly interpretable through their regression coefficients. Using the regression coefficients, operational regression models such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan 1994) can display the individual effect of each element of the present meteorology on the final prediction. With care, this also allows interpretation of the relative utility of different pieces of meteorological information in predicting the forecast phenomenon of interest, in this case, tropical cyclone intensity (e.g., Jones et al. 2006). Direct inspection of the parameters is equally insightful for other types of regression, such as in multivariate logistic regression (LR) for probabilistic forecasts (e.g., Bremnes 2004). While direct interpretability is an attractive quality of regression models, the parametric nature of them and like algorithms imposes assumptions on the relationship between the predictors and the predictand or between predictors themselves when such relationships may not be accurate or even known or physically understood (Wilks 2011). Linear and logistic regression, for example, both impose a fundamentally linear predictor–predictand relationship and treat predictors independently, not directly accounting for the covariance between multiple predictors and their joint relationship with the predictand. While imposing these restrictions can actually be helpful when they are physically valid, predictive performance degrades when these imposed assumptions are invalid.

Especially when the physical relationships are not known or well quantified, it is often attractive to employ an algorithm that does not impose such assumptions. One such example is the random forest (RF) technique (Breiman 2001). RFs have been used for many different applications in NWP, including but not limited to prediction of storm-type classification, turbulence, cloud ceiling and visibility, convective initiation, and hail size (e.g., Williams 2014; Herman and Schumacher 2016b; Ahijevych et al. 2016; Gagne et al. 2017; McGovern et al. 2017). Though the algorithm is more general, the inner workings of an RF may be diagnosed, like with regression coefficients for LR, primarily by means of feature importances (FIs) to be used and discussed in more detail in this study. While these have already been used to assess RF NWP models in some past studies (e.g., Gagne et al. 2014; Herman and Schumacher 2016b), in using locally extreme precipitation forecasting as an example, we will demonstrate here that they can be used to understand spatiotemporal relationships, as well as relationships across atmospheric fields in predicting the phenomenon of interest—even when the number of predictors grows large, the event becomes rare, and algorithmic steps that complicate the relationship between the predictor inputs and reality are performed.

Herman and Schumacher (2018) expanded upon these prior studies using machine learning for NWP in a variety of ways. While there have been limited prior studies using machine learning to explicitly investigate very rare events (e.g., Marzban and Stumpf 1996; Marzban and Witt 2001) and some prior studies constructing statistical models for quantitative precipitation forecast (QPF; e.g., Hall et al. 1999; Sloughter et al. 2007; Whan and Schmeits 2018, manuscript submitted to *Mon. Wea. Rev.*), there has been little published work to date combining both facets. Herman and Schumacher (2018) were among the first to do so, training statistical models to forecast locally extreme precipitation across the contiguous United States (CONUS) in the medium range. The CONUS-wide gridded scope of the models trained therein is also uncommon among machine learning models, which are often trained for points (e.g., Herman and Schumacher 2016b) or over a limited domain (e.g., Gneiting et al. 2005). Furthermore, the scope of predictors was very large, with thousands of predictors capturing the spatiotemporal environmental characteristics of the forecast point during the accumulation period. Many different sensitivity experiments were performed, and the performance of the model forecasts was evaluated in detail from both the perspective of forecast skill and reliability. Overall, forecasts were found to add both considerable skill and reliability across all of the CONUS, compared with both climatology and the raw forecasts of the global ensemble from which the model predictors were derived. However, the study did not investigate the internals of these models: that is, how to visualize what they are doing to get from their input to their output, and what these algorithms and models reveal about the prediction of locally extreme precipitation events overall.

Using the regression and tree-based models of Herman and Schumacher (2018), this "dendrological" study investigates the details of the fitted trees, as well as the regression models. We illustrate how models based on seemingly abstract and complex algorithms and techniques can, with modest effort, be readily interpreted and understood. It is shown that not only can these models yield more skillfully verifying forecasts than raw dynamical model output or forecasts derived from simpler, more traditional postprocessing approaches, but they can also provide both statistical and physical insights into why they behave as they do, as well as insight into the deficiencies, errors, and limitations of the dynamical model predictors on which they are based. In this study, examination of the Herman and Schumacher (2018) models sheds insights onto how a global, convection-parameterized dynamical ensemble behaves in forecasting extreme precipitation events across the hydrometeorologically diverse regions of the CONUS, and on what statistical corrections can be made to improve forecasts thereof. Section 2 briefly summarizes the methods of Herman and Schumacher (2018) to describe the underpinnings of the models evaluated in this study and how they were derived. Section 3 describes how the models will be visualized and interpreted in this study. Sections 4, 5, and 6 present results, respectively, for PCA diagnostics, RF models, and LR models. Section 7 concludes with a synthesis of the findings and a discussion of their implications.

2. Data and methods summary

What follows is an abbreviated description of the full data and methods of Herman and Schumacher (2018), highlighting the aspects that are critical for proper interpretation of the results presented herein. The interested reader is encouraged to review the full methods of that study for a more complete discussion of the mathematical underpinnings of the algorithms, the justification of choices made, and the sensitivity experiments performed therein.

The models evaluated in this study are trained to forecast locally extreme precipitation across the CONUS for 24-h precipitation accumulations, quantified with respect to average recurrence interval (ARI) exceedances. In particular, models are trained to issue probabilistic forecasts for exceedances of 1- and 10-yr ARIs within a ~0.5° × 0.5° spatial domain during a 24-h 1200–1200 UTC accumulation interval. Forecasts are made for two different forecast lead times comprising the 36–60- and 60–84-h periods—denoted, respectively, day 2 and day 3—with separate models trained for each period. Unique models are also trained for each of eight different geographic regions of the CONUS, as depicted in Fig. 1. Here, the CONUS has been partitioned to produce cohesive regions with some hydrometeorological homogeneity with particular regard to similar magnitudes of extreme precipitation, similar diurnal and seasonal precipitation climatologies, and similar storm types and precipitation processes associated with extreme precipitation.

Dynamical model data used for training the statistical models in this study come from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) dataset. The GEFS/R is an 11-member ensemble with T254L42 resolution—which corresponds to an effective horizontal grid spacing of ~55 km at 40° latitude—initialized once daily at 0000 UTC back to December 1984. Forecast fields evaluated in this study are archived on a grid with ~0.5° horizontal spacing. For day 2 models, forecast fields use 3-h temporal resolution, while 6-h resolution is used for day 3 models. Trained models discussed in this study are based on the ensemble median of a core set of nine atmospheric fields: accumulated precipitation (APCP), surface-based convective available potential energy (CAPE) and 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). Models are trained using daily forecasts spanning from January 2003 through August 2013.

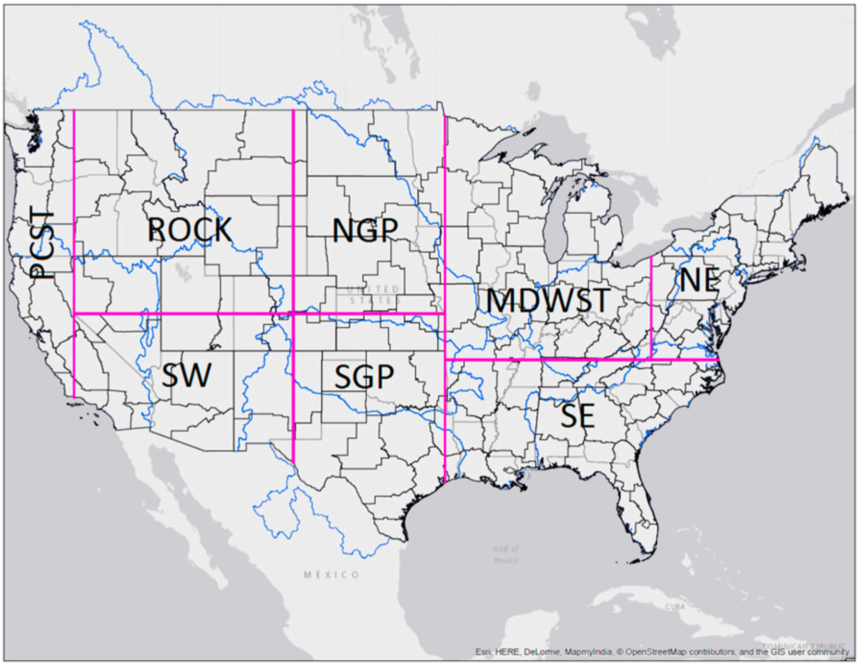
PCST
ROCK
NGP
MDWST
NE
SW
SGP
SE
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

FIG. 1. Map depicting the regional partitioning of the CONUS used in this study and the labels ascribed to each region. Adapted from Herman and Schumacher (2018).

The National Centers for Environmental Prediction (NCEP) stage IV precipitation analysis product (Lin and Mitchell 2005) has been created daily in an operational capacity since December 2001. Stage IV provides 24-h analyses over the CONUS on a ~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), its analysis quality, resolution—allowing relatively accurate quantification of very heavy precipitation—and data record length make it preferable to other precipitation analysis products and is therefore used as the precipitation "truth" for this study.

The return period thresholds (RPTs) associated with the 1- and 10-yr ARIs 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 they were available at the time this research began. For five northwestern states—Washington, Oregon, Idaho, Montana, and Wyoming—updated thresholds are not available, and derived Atlas 2 threshold estimates are used instead (Miller et al. 1973; Herman and Schumacher 2016a). In the Northeast—New York, Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island—and Texas, both of which did not have Atlas 14 threshold estimates at the time research commenced but have either since received an update or have an update in progress, Technical Paper 40 (TP-40; Hershfield 1961) estimates are used. Everywhere else uses the Atlas 14 RPT estimates.

Generating predictors by taking GEFS/R forecast values from nine different fields every 3 or 6 h over a 24-h forecast period at every grid point within ~2° of the forecast point yields thousands of model predictors. In addition to the large quantity, they are also highly correlated—spatially, temporally, and across variables. With millions of training examples and thousands of predictors, the forecast problem can become computationally intractable, and the correlated variables can result in overfitting. To address these issues, use of a preprocessing step whereby the model predictors undergo dimensionality reduction via principal components analysis (PCA) is explored. This creates a small set of uncorrelated predictors that explain the signal in the forecast data and give insight into the regional modes of atmospheric variability as depicted in the GEFS/R model, while leaving the noise in withheld lower-order principal components (PCs). While PCA has been mostly applied in the atmospheric sciences for

TABLE 1. Summary of the models trained in this study and the corresponding names designated to the models. An "×" indicates the process is performed or the information is used; a lack of one indicates the opposite. MEDIAN corresponds to the ensemble median. Horizontal radius is listed in grid boxes from forecast point; time step 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. Models apply to all eight forecast regions and have both day 2 and day 3 versions.

Model name	CTL_NPCA	CTL_PCA	CTL_LR
Algorithm	RF	RF	LR
PCA preprocessed		×	×
Ensemble information	MEDIAN	MEDIAN	MEDIAN
Horizontal radius	4	4	4
Time step	3/6	3/6	3/6

identifying spatial patterns at the largest scales (e.g., Thompson and Wallace 1998; Wheeler and Hendon 2004), flavors of PCA have been successfully applied to identify smaller synoptic and mesoscale features as well (e.g., Mercer et al. 2012; Peters and Schumacher 2014).

Herman and Schumacher (2018) performed a wide array of sensitivity experiments, exploring model predictive performance as a function of predictor temporal resolution, spatial extent, inclusion or exclusion of different atmospheric fields, use of ensemble information, algorithmic parameters, and choice of model algorithm. To manage the scope of this study's analysis, only results as a function of the last of these is presented. Much like the skill results presented in Herman and Schumacher (2018), general physical findings are found not to vary appreciably as a function of any of these unshown dimensions of variability. Three models of Herman and Schumacher (2018) specifically are evaluated in depth in this study: 1) the CTL_NPCA model using random forests and no PCA dimensionality reduction; 2) the CTL_PCA model using random forests with preprocessing using PCA dimensionality reduction; and 3) the CTL_LR model using logistic regression and also using PCA preprocessing. Table 1 provides a summary comparison of these three models for reference.

Random forests (Breiman 2001) are in essence an ensemble of decision trees, whereby each tree of the forest makes an individual prediction of the predictand outcome; the relative frequencies of each possible outcome in the ensemble of trees are then used to make a single probabilistic forecast. Decision trees are explained in mathematical depth in Herman and Schumacher (2018); an alternative way to conceptualize them begins with a many dimensional predictor phase space, where each predictor has a unique dimension. Beginning with an unpartitioned phase space (tree root), a decision tree makes successive splits along axes of this space, partitioning it into increasingly many smaller subspaces (splits) and then assigning predictions to each subspace (leaves). An RF creates many different similarly plausible partitions of the subspace, and a forecast is determined by the subspace labels associated with the given point in predictor space.

Logistic regression is an implementation of the generalized linear model, designed for binary predictions and classification more generally where the predictand is constrained to be either one outcome or another, rather than over a continuous space as with linear regression (Wilks 2011). Like with linear regression, logistic regression uses as its input a linear combination of the predictors. The difference arises in the use of the link function. For linear regression, the link is the identity function; that is, the prediction is the aforementioned linear combination of the predictors. In the case of logistic regression, the predictor–predictand link is made through use of the logit function instead (Wilks 2011). In particular, the model output in multinomial logistic regression—the probability of each event class—is given by use of a generalization of the logistic function:

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

where k is the event class, $\mathbf{x}$ is the predictor vector, and $\mathbf{w}$ is the vector of regression coefficients. Note that separate coefficient vectors are computed for each event class.

3. Methods: Model properties and assessment

One of the most powerful aspects of machine learning algorithms—and RFs in particular—is finding patterns in the supplied training data. Because of the extent and diversity of the data supplied in these experiments, the RFs trained for this study have the theoretical capability of diagnosing and automatically correcting for various kinds of GEFS/R model biases. In particular, context-dependent quantitative biases, such as GEFS/R QPF being systematically too high or too low, may be diagnosed; spatial displacement biases in the placement of extreme precipitation features may be diagnosed; and temporal biases in the initiation or progression of extreme precipitation features may also be diagnosed to some extent. These can be at least partially visualized through RF FIs. The most intuitive way to conceptualize their quantitative significance is by the number of splits based on the feature summed over the forest, with each split weighted in proportion to the number of training samples encountering the split so that a split at the root node is considered much more important than a split deep into a tree (Friedman 2001). Values are normalized so that the sum of all importances is one; an importance of one then indicates that all decision nodes in every tree of the forest split on the corresponding feature, while an importance of zero indicates that no decision node splits based on that feature. Importances are produced for each input feature; without PCA preprocessing, this means that an individual importance value is produced for each forecast point-relative location–forecast time–atmospheric field combination. In many cases, it is convenient to present importances summed over one or more of these dimensions for a summary perspective of the model output. When PCA preprocessing is performed, the model output is instead importances of individual PCs in predicting ARI exceedances. FIs calculated in this way—often termed the "Gini importance"—are only one method of providing a summary representation of an RF (Strobl et al. 2007). In the leading alternative method, the so-called "permutation accuracy importance" approach (Strobl et al. 2008), for each predictive feature, the feature value for each sample used to construct a given tree is permuted to a different sample's value. Importance is determined by the decline in the model's predictive performance when replacing the true values with the permuted ones. This is calculated individually for each tree and then averaged over the entire forest. While this approach has some advantages over other approaches (e.g., Breiman 2001; Strobl et al. 2007, 2008), the "Gini importance" measure is used for this study for consistency with past studies in the field (e.g., Herman and Schumacher 2016b; Gagne et al. 2017; Whan and Schmeits 2018, manuscript submitted to *Mon. Wea. Rev.*) and computational simplicity (Pedregosa et al. 2011).

One of the main advantages of LR is its interpretability; through the regression coefficients, there is a direct, concrete connection between the predictors and the forecast predictand. And although the regression in the CTL_LR model is performed on the principal components and not the native atmospheric variables, the relationship to the native features may be readily backed out through the PCA loadings matrix $\mathbf{L}$:

$$
\begin{aligned}
\mathbf{x}^{\mathrm{T}}\mathbf{w}_k &= w_{k,1}\mathrm{PC1} + w_{k,2}\mathrm{PC2} + w_{k,3}\mathrm{PC3} + \cdots + w_{k,R}\mathrm{PC}R \\
&= w_{k,1}\left(\sum_{m=1}^{M}\mathbf{L}_{1,m}\mathbf{F}_m\right) + w_{k,2}\left(\sum_{m=1}^{M}\mathbf{L}_{2,m}\mathbf{F}_m\right) \\
&\quad + w_{k,3}\left(\sum_{m=1}^{M}\mathbf{L}_{3,m}\mathbf{F}_m\right) + \cdots + w_{k,R}\left(\sum_{m=1}^{M}\mathbf{L}_{R,m}\mathbf{F}_m\right),
\end{aligned}
\tag{2}
$$

which yields

$$
\begin{bmatrix} F_1 & F_2 & F_3 & \cdots & F_M \end{bmatrix} = \begin{bmatrix} L_{1,1} & L_{1,2} & L_{1,3} & \cdots & L_{1,R} \\ L_{2,1} & L_{2,2} & L_{2,3} & \cdots & L_{2,R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_{M,1} & L_{M,2} & L_{M,3} & \cdots & L_{M,R} \end{bmatrix} \times \begin{bmatrix} w_{k,1} \\ w_{k,2} \\ \vdots \\ w_{k,R} \end{bmatrix},
\tag{3}
$$

where R is the number of retained PCs, M is the number of native features, k is the event class, $\mathbf{F}$ is the vector of native features, and $\mathbf{w}$ is the vector of regression coefficients.

Both algorithms have their advantages, disadvantages, and caveats in interpretation. As noted above, LR has the advantage of a direct quantitative link between any given predictor of interest and the predictand. RF FIs, in contrast, give only an "importance" number, which gives no indication of the sign or magnitude of the predictor in order to correspond with event observance. While it can be executed, the task of manually inspecting the value of every node split based on the predictor is cumbersome, and it is difficult to draw general conclusions due to the deeply layered subspaces involved. However, RFs do have major advantages over LR in interpretation as well. As a linear model, LR coefficients are constrained to apply globally, but this is often not an appropriate constraint. Some predictors may only become important when other conditions are satisfied—for example, CAPE might only be important when there is a lifting mechanism to release the instability—rendering them insignificant in most cases but very important under select circumstances. In LR, where the coefficient applies uniformly regardless of the circumstances, the regression coefficient would necessarily be small, while the RF FI for the same predictor could be relatively large by harnessing the predictive

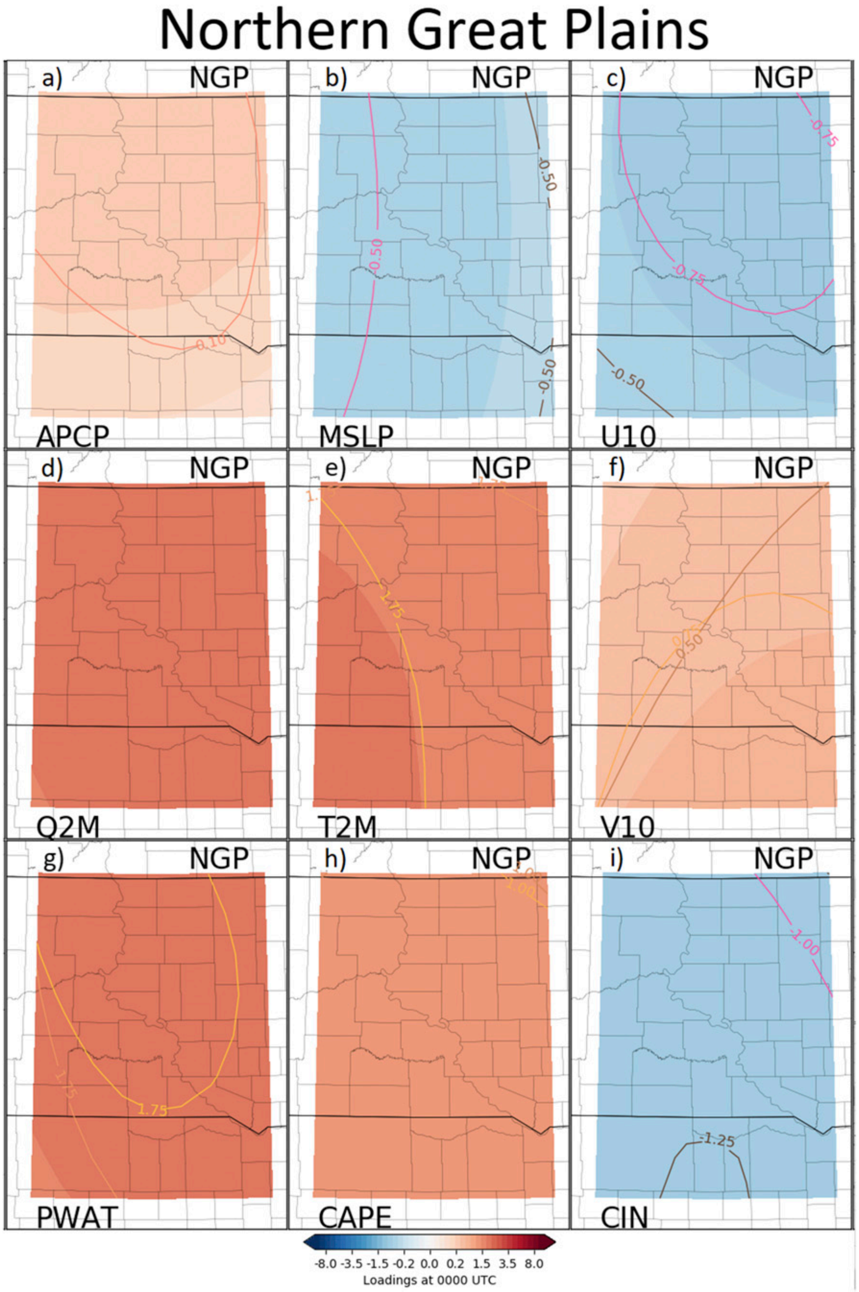

FIG. 2. PC1 loadings for the NGP region. (a)–(i) Loadings associated with the APCP, MSLP, U10, Q2M, T2M, V10, PWAT, CAPE, and CIN fields, respectively. Filled contours depict loadings for forecast values at 0000 UTC during the forecast period (forecast hour 48), with reds indicating positive loadings and blues negative loadings; magenta and yellow contours indicate negative and positive loadings, respectively, for 1500 UTC during the period (forecast hour 39), while brown and beige contours depict negative and positive loadings for 0900 UTC during the forecast period (forecast hour 57). Darker colors indicate larger values and, accordingly, a stronger relationship with the principal component as indicated in the figure color bar.

utility of the variable within a particular subspace of the larger predictor space. RFs also handle correlated predictors better than LR. In regression problems, when one predictor is highly correlated with another, one is liable to have a situation whereby the ‘‘weight’’ is disproportionately allocated to one predictor over the other, giving the false appearance that one variable is highly predictive while the other is not. In RFs, with two highly correlated predictors that are thus approximately equally predictive, node splits will occur essentially randomly between one and the other, and the RF FIs thus have a tendency to balance approximately evenly (Gagne 2016). This problem of LR is greatly alleviated in the CTL_LR model by using PCs as the predictors, which are necessarily constructed to be orthogonal to one another. However, analyzing these different algorithmic formulations in tandem enables capturing a more complete picture of the extreme precipitation forecast problem.

4. Results: GEFS/R principal components analysis

Inspection of the leading mode of atmospheric variability—PC1, the component that explains the most variance between different days or model initializations—in Figs. 2 and 3 for the northern Great Plains (NGP) and Pacific coast (PCST) regions, respectively, and for the remaining regions in the online supplement, reveals that the leading mode in each region quite apparently relates to the seasonal cycle. However, the precise nature of that seasonal cycle varies by region. Like colors across subpanels in these figures indicate that atmospheric fields covary together for the region's displayed PC, while contrasting colors indicate one variable is anomalously high while the other is low. Deeper reds associate positively with the PC, with blues associating negatively; lighter colors indicate that the given predictor does not relate as strongly with the PC. Spatial color inhomogeneities within a subpanel suggest the PC is associated with a spatial gradient in the field, while loadings changing throughout the forecast period—shown via comparison of the unfilled contours—are indicative of some degree of regime change. By happenstance, positive values of PC1 in all regions are compared with a summer signal, while negative values are associated with a winter signal. In all regions, the summer signal is associated at all times of day with high surface temperature and moisture (e.g., Figs. 2d,e), higher PWAT and CAPE (e.g., Figs. 2g,h), and lower MSLP and CIN (e.g., Figs. 2b,i). In almost every region, the warm-season signal (positive PC1) is weakly associated with APCP for the region (e.g., Fig. 2a); in other words, this states that the warm season is also the wet season in most regions of the CONUS. However, in the PCST region, precipitation is predominantly received during the cool season (Herman and Schumacher 2016a), and this is reflected by negative loadings for the APCP field seen in Fig. 3a. The primary regional differences between the seasonal cycle and reflected in the PC1 loadings is seen in the wind fields (Figs. 2c,f and 3c,f). In most regions, including NGP, the warm season is associated with anomalous southeasterly flow at low levels, as evidenced by positive V10 loadings (Fig. 2f) and negative U10 loadings (Fig. 2c). However, this is not true of the western regions; PCST, like APCP, exhibits the opposite behavior to the eastern regions in the wind fields, with a warm season characterized by anomalous northwesterly flow (Figs. 3c,f). The strength of association with PC1 also varies between atmospheric fields. The seasonal cycle, at least as reflected in PC1, is predominantly a thermodynamic and moisture signal; this is seen by observing larger loading magnitudes with fields such as Q2M, T2M, and PWAT, compared with APCP and other fields (cf. Figs. 2d,e,g and 2a–c,f,i).

In one sense, the seasonal cycle, and thus PC1, is rather trivial—it is already largely known and understood. It would be possible to train these models with deseasonalized predictors and an additional predictor(s) to represent location in the seasonal cycle, and this prospect is worthy of further investigation in future work. But this could appreciably harm predictive performance of the model; in many instances, a certain quantity of a precipitation ingredient such as precipitable water or CAPE (e.g., 35 mm or 1500 J kg^{-1}, respectively) is necessary to generate locally extreme precipitation-producing storms. By instead supplying deseasonalized predictors, these physical thresholds, which may be climatologically much more likely in one season than another, are severed from the numerical values of the model predictors. This forces the model to, in essence, relearn the seasonal cycle via a combination of the seasonal indicator predictor and deseasonalized atmospheric predictors, in addition to all of the other relationships it must diagnose, placing an extra burden on model training. This would likely sacrifice predictive accuracy of the trained models, perhaps with the gain of a more physically insightful PC1.

PC2—the leading mode of atmospheric variability at a point aside from the seasonal cycle—is depicted for the NGP and PCST regions in Figs. 4 and 5; PC2 loadings for other regions may be found in the online supplement. While PC1s were largely similar between the regions, there are substantial regional differences between the PC2 loadings. Generally, while PC1 is predominantly a thermodynamic signal, many PC2s are predominantly a kinematic signal, with the largest loading magnitudes typically seen in U10 and V10. Furthermore, while PC1 loadings had little temporal dependence, for PC2 and beyond, loadings changing sign or magnitude across the forecast period are commonplace (e.g., Figs. 4a,b,i). One notable commonality is that in many regions, PC2 shares at least some characteristics one might expect associated with frontal passage, including rapid changes in meridional winds [e.g., southeast (SE), southwest (SW), and northeast (NE) regions—see online supplement], pressure

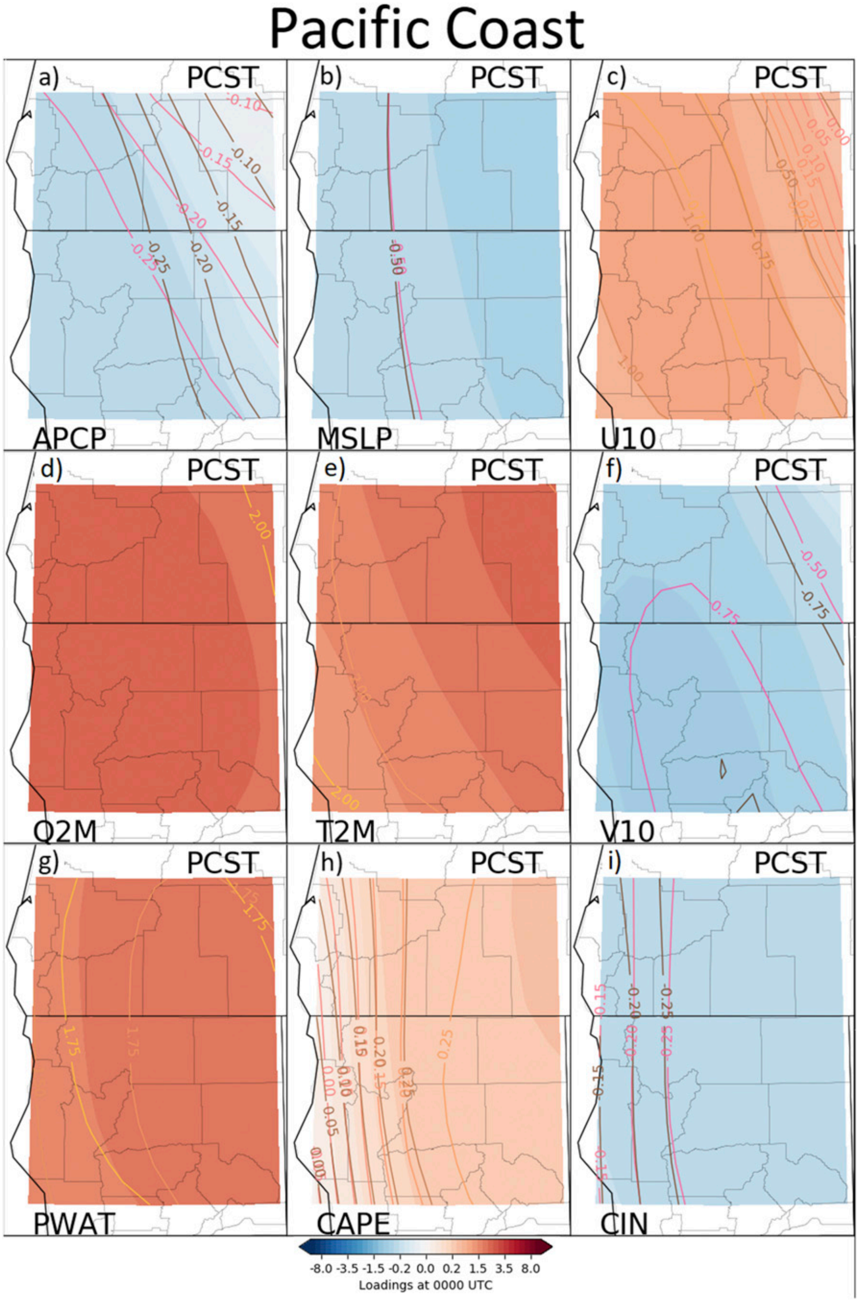

FIG. 3. As in Fig. 2, but for the PCST region.

falls (e.g., Fig. 4b), precipitation and moisture changes (e.g., Figs. 4d,g), and even instability "advection" [e.g., southern Great Plains (SGP)]. In the PCST region (Fig. 5), where fronts are thermodynamically weak compared with other regions, they govern a smaller portion of atmospheric variability in the region and are not associated with PC2. The signal looks somewhat atmospheric river–like, with heavy precipitation (Fig. 5a), column-integrated moisture advecting in from the southwest with strong low-level southerly flow (Figs. 5f,g), and low pressure and temperature (Figs. 5b,e), at least when compared with the warm season. Again, though, some loadings do not appear entirely consistent with this interpretation (e.g., Fig. 5c). None of the PC2 loadings appear to have a direct physical interpretation that clearly matches with every aspect portrayed by the PC, a known drawback imposed by the combined orthogonality and maximum variance limitations imposed in the PCA formulation (e.g., Richman 1986).

5. Results: RF diagnostics

Associated with an RF is a single FI for each predictor. When no dimensionality reduction is performed in advance, there are thousands of GEFS/R predictors; each predictor is associated with a particular atmospheric field, forecast hour, latitude, and longitude. In addition, there is a unique RF for each of the eight regions and each of the two forecast periods. Effectively visualizing and interpreting all of these FIs can be difficult. To manage the visualization task, RF FIs are first presented by considering only one dimension of predictor variability at a time. For example, FIs are considered as a function of the atmospheric field associated with the predictor, without regard to the hour or forecast point-relative location of the predictor. FIs are then considered by grouping all predictors with the same forecast hour, and last by grouping predictors with the same forecast point-relative location. This allows tractable visualization of a summary of the FI output of the GEFS/R and helps identify areas for more detailed analysis of a subset of "raw" (single predictor) FIs presented in the second half of this section. For the interested reader, the full set of RF FIs is included in an online supplement to this paper.

GEFS/R QPF, or APCP, is reliably identified as one of the most predictive atmospheric fields for observed extreme precipitation based on RF FIs summed over space and time for each region of the day 2 version of the CTL_NPCA model (Fig. 6). This indicates that the dynamical model, in this case the GEFS/R, has some skill in directly simulating extreme precipitation. However, the extent of model APCP being predictive over other ingredients-based fields varies substantially by region. In the PCST region, where extreme precipitation events are predominantly driven by atmospheric rivers and other large-scale systems advecting moisture over orography (e.g., Rutz et al. 2014; Herman and Schumacher 2016a), a convection-parameterized model such as the GEFS/R is able to adequately simulate the largely stratiform precipitation processes. This is reflected in the RF FIs shown in Fig. 6e; the model APCP, which adequately captures the processes involved in producing most precipitation events in the region, has a total FI of approximately 50% of the total, more than 5 times that of any other field. In other regions that feature a mix of synoptic and convective events, such as Rocky Mountains (ROCK), NE, and SE (respectively, Figs. 6a,d,h), APCP is still by far the most important atmospheric field in predicting observed APCP, but to a much smaller degree than in the PCST region, with values in the 0.25–0.4 range. In the regions where extreme precipitation events are most driven by convective-scale processes unresolvable by the GEFS/R and that correspondingly have the poorest verifying raw QPFs in predicting extremes (Herman and Schumacher 2016a), such as NGP and Midwest (MDWST) (Figs. 6b,c), model APCP is not even the most important atmospheric field in predicting ARI exceedances. While still somewhat important, with aggregate RF FIs of approximately 0.18, APCP is identified as less predictive than PWAT in these two regions, with PWAT FIs in the 0.25–0.35 range. One physical explanation is that where the GEFS/R is poor at predicting extreme precipitation events by virtue of an inability to resolve the responsible processes, ingredients such as column-integrated moisture become more useful predictive tools. PWAT remains a valuable predictor in other regions as well, with greater importances also observed in the ROCK, NE, SGP, and SE regions (Figs. 6a,d,g,h). In one region, the SW (Fig. 6f), surface moisture (Q2M) was considered more predictive than column-integrated moisture (PWAT), but this was not generally the case. In most cases, CAPE and CIN were the least predictive fields among those examined, but the SW region (Fig. 6f) was again a considerable anomaly, with CAPE and CIN being, respectively, the second and third most important fields, and CAPE FIs nearly equal to those of APCP. Regional RF FIs at day 3 look largely similar to the day 2 RF FIs, but some minor differences can be discerned. The APCP RF FIs are slightly lower in many of the regions, particularly in the eastern regions (Figs. 6d,h). In general, "ingredients"—fields other than the direct APCP from the GEFS/R—are relied on somewhat more at day 3, compared with day 2.

Time series of RF FIs shed insight into which times forecast guidance provides the most useful predictive information for the quantity of interest, in this case ARI exceedances, and can help identify systematic biases in the parent model's diurnal climatology of relevant processes, such as convective initiation. They can also provide insight into when particular information is of value—whether the information is useful as a precursor or concurrent with the actual precipitation. Every region exhibits broadly similar FI time series when aggregated over all variables (Fig. 7, red and blue lines), with importance minima at both 1200 UTC times—the beginning and end of the forecast period—and a maximum during the middle of the day. A combination of two reasons likely explains this pattern. First, the middle of the day, in the afternoon and evening hours, is typically the most convectively active and is the period in which precipitation and heavy precipitation are most frequently observed (e.g., Stevenson and Schumacher 2014; Herman and Schumacher 2016a). Second, it is also, somewhat coincidentally, the middle of the forecast period, and thus forecast values at this time can be more

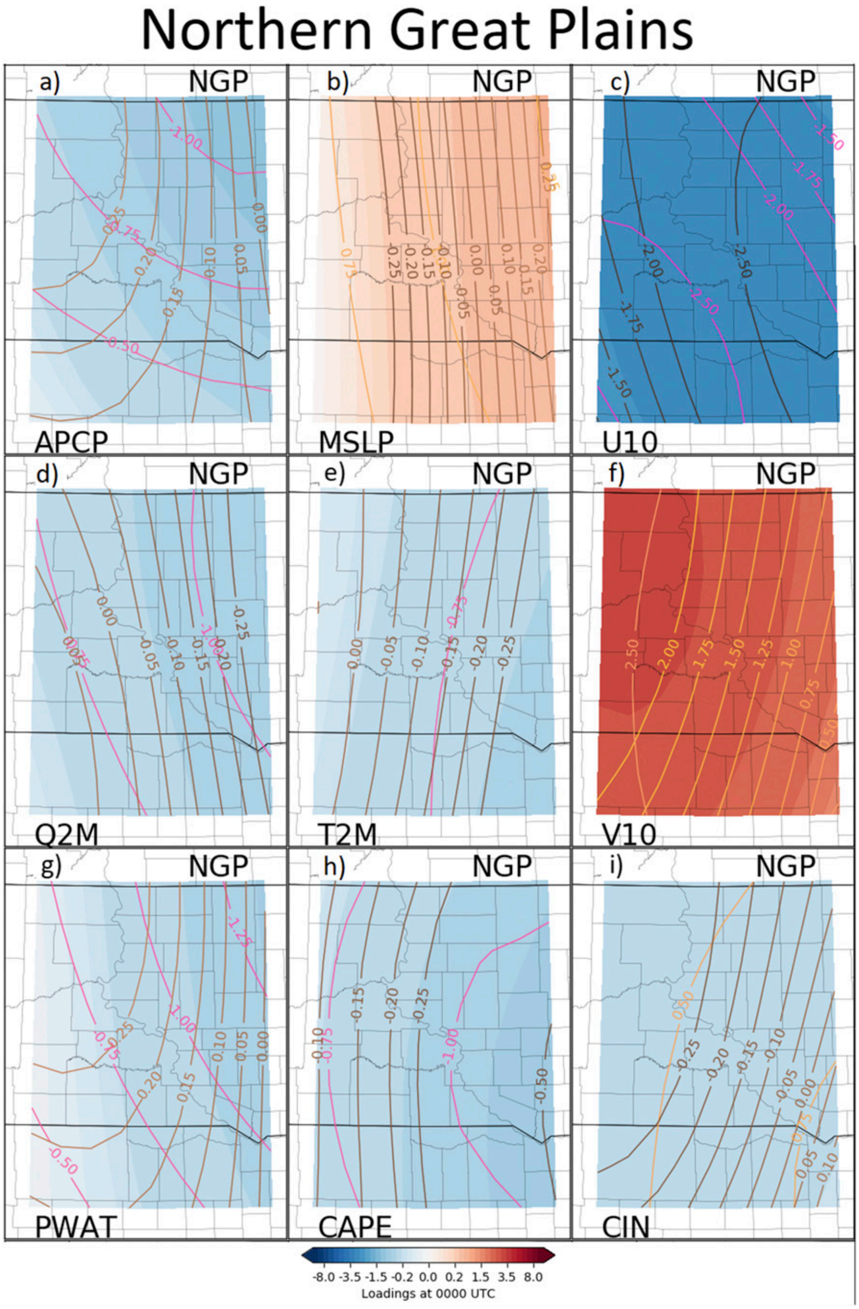

FIG. 4. As in Fig. 2, but for PC2.

representative of the period as a whole. In most regions, the difference between the minimum and maximum importance values for a given forecast time spans approximately a factor of 2. There is also more variability in the time-dependent FIs, comparing, for example, the relative width of the red- and cyan-shaded regions in the panels of Fig. 7 with the error bars of Fig. 6. Perhaps the most important finding is that the FI time series partially reflect the diurnal climatology of extreme precipitation events in each region. FIs are higher later in the forecast period in regions where extreme precipitation events tend to occur in the evening and overnight, such as the MDWST and SGP (Figs. 7b,c,g), while regions where events tend to be more in the afternoon hours, such as the NE and SE (Figs. 7d,h), have a peak at 0000 UTC and a significant dropoff in importance by 0600 UTC. While this is seen in the time series with all fields aggregated, it is especially pronounced

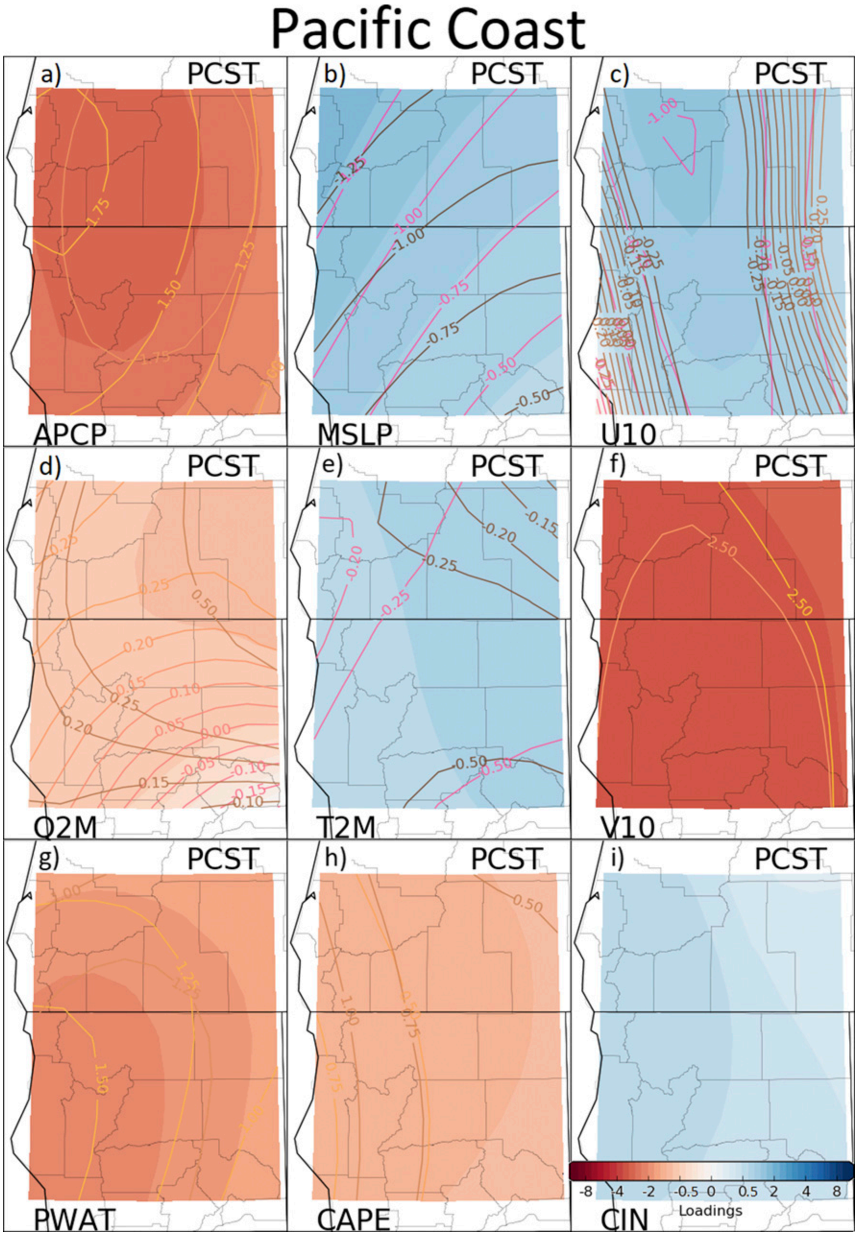

FIG. 5. As in Fig. 3, but for PC2.

when considering only the APCP FIs (Fig. 7, purple and maroon lines). While the APCP FIs follow the diurnal precipitation climatology specific to the forecast region, PWAT FIs maximize prior to the maximum APCP FIs, particularly in regions where PWAT is found to be predictive (e.g., Figs. 7b,c,g), sensibly indicating that the column moisture of the environments in which storms form is an important property for predicting locally extreme precipitation.

Compared with the time series of Fig. 7, more stark regional contrasts are observed for FIs compared in space (Fig. 8). As would be naively assumed, some regions have an importance maximum near the forecast point, with decreasing importance with increasing distance from the forecast location. This is broadly true of the ROCK, PCST, SW, SGP, and SE regions (Figs. 8a,e–h). The other three regions—NGP, MDWST, and NE (Figs. 8b–d)—have an importance maximum well downstream of the forecast point. This summary view does not provide insight into the precise physical reasons why this may be the case; possible causes include a combination of precipitation features moving

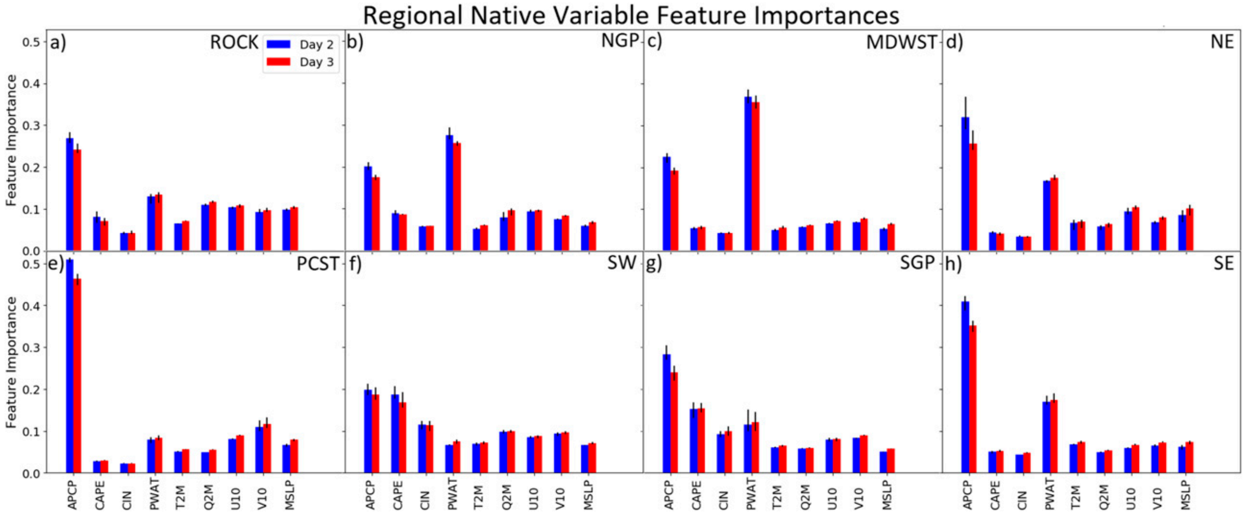

FIG. 6. Regional comparison of the summed RF FIs for the different atmospheric fields used in the CTL_NPCA model, summed over the time and two spatial dimensions. The blue bars correspond to the mean summed FIs of the four models trained via cross validation for the day 2 version of the model; red bars correspond to the day 3 model version. Error bars indicate the minimum and maximum cross-validation summed FIs. (a)–(h) ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions, respectively.

too quickly, progged systems developing too far downstream, or that the downstream environment is simply better predicted than the environment in which the extreme precipitation events occur, and thus serves as a better predictor than the fields collocated with the forecast point. More investigation into possible reasons will be discussed below. Several other interesting regional differences may be noted. Some regions, such as PCST and SW (Figs. 8e,f), have a highly concentrated spatial maximum—with differences in importances between forecast points spanning nearly an order of magnitude—meaning that information from a particular location is much more predictive than surrounding areas. This likely indicates both increased persistence and consistency of model biases in these regions, as well as enhanced predictability overall, consistent with the higher forecast skill in these regions (Herman and Schumacher 2018). It also suggests that the RF is likely tracking specific simulated GEFS/R precipitation features in these regions, as opposed to just predicting based on a general characterization of the environment in which storms might form, which would yield more spatially homogeneous FIs. The five aforementioned regions with a maximum FI point near the forecast point also do not all have these two points exactly collocated. In the PCST region (Fig. 8e), the point of maximum importance is displaced slightly to the south and west of the forecast point. This is true to some degree in the SW and SE regions as well (Figs. 8f,h). Meanwhile, a slight north and particularly west displacement is seen in the SGP region (Fig. 8g). These displacements may indicate persistent biases in the portrayal of extreme precipitation elements and/or the ingredients responsible for them. In the ROCK, SGP, and SE regions, a secondary maximum well downstream of the forecast point is observed in a pattern resembling that of the other northern regions. In the regions that do have a downstream maximum, either primary or secondary, the more western regions—ROCK, NGP, and SGP—have the maximum also displaced well to the north (and east), while the regions farther east, such as MDWST and NE, have the maximum to the south.

Raw FIs for the APCP field in the day 2 version of the CTL_NPCA model (Fig. 9) reveal that, consistent with Fig. 7, APCP importances increase to a daytime or evening maximum with importance minima at 1200 UTC, with the strength of the cycle varying by region. Because the accumulation interval lies outside the forecast period for the front-end 1200 UTC time, the importance is identified as the lowest there, compared even with the 0600–1200 UTC QPF at the end of the forecast period. Correspondingly, in some regions (e.g., Figs. 9b1,c1), there is a lack of a clear, cohesive precipitation feature—as represented by an importance maximum—at the beginning of the forecast period. However, at this time or subsequent to it, a clear importance maximum in the precipitation field emerges in each region and can be seen to track from west to east across the forecast point-relative domain throughout the forecast period, tracking the typical progression of precipitation systems with the mean upper-level flow. At the beginning of the forecast period, FI maxima (Fig. 9, column 1) are located 1–2 grid points west of the forecast point, while by the end (column 5), they are

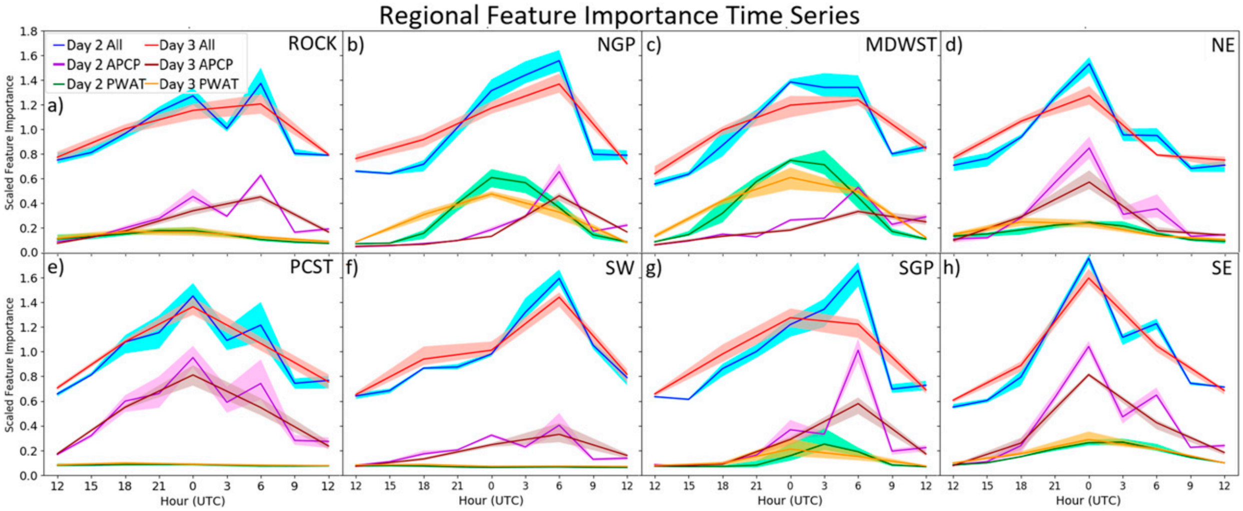

FIG. 7. Regional comparison of the summed RF FI time series in the CTL_NPCA model. Blue and red lines depict, respectively, the day 2 and 3 versions of the model, summed over both spatial dimensions and all atmospheric fields. Values have been renormalized based on the number of time periods for the version of the model so that the a priori expected importance for each time is unity. The purple and maroon lines depict the day 2 and day 3 FI time series for only the APCP predictors, summed over the two spatial dimensions. Green and yellow lines are as with the purple and maroon lines, respectively, but for the PWAT FIs. The same normalization is applied to these time series as well, leaving a priori expected summed FIs of unity divided by the number of atmospheric fields (nine). Shading about each line indicates the range of values obtained through the fourfolds of cross validation, with the lines themselves representing mean values of the fourfolds. (a)–(h) ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions, respectively.

located anywhere from 0 to 3 grid points displaced to the east, with meridional alignment in the PCST region (Fig. 9a) and far eastern displacement on the five easternmost regions (e.g., Figs. 9b–d). This may be diagnosing regional climatological differences in the progression speed of extreme precipitation-producing systems, which may remain relatively stationary over the complex terrain of PCST, while moving quickly over the flatter terrain farther east. But another important factor that it may be identifying is model biases in the progression of extreme precipitation systems; it may be noting that GEFS/R systematically moves systems in the east too quickly and systems in PCST perhaps too slowly, resulting in APCP well downstream of the forecast point being predictive of extreme precipitation in the eastern regions in a way that it is not in the western regions. More investigation is required to diagnose the extent to which each of these factors is in play in yielding this end diagnosis. Of further interest are the different progressions of FI maxima across different regions. In the five regions east of the Rocky Mountains—NGP, MDWST, NE, SGP, and SE (e.g., Figs. 9b–d)—a clear southwest–northeast progression is seen and is particularly pronounced in the SGP region. The regions meridionally aligned with the Rocky Mountains, ROCK and SW, have little latitudinal variation with time, though a slight southwest–northeast is observed in ROCK and a slight northwest–southeast observed in SW (see online supplement). PCST, in contrast to most of the other regions, has a clear northwest–southeast temporal FI progression (cf. Figs. 9a1 and 9a5). These progressions are consistent with the typical synoptic flow of locally extreme precipitation environments of these regions. The southward progression of postlandfall atmospheric rivers warrants further investigation but is consistent with some previous studies (e.g., Ralph et al. 2011), and the southwest–northeast progression in the northeast is consistent with both tropical cyclones, which are almost always progressing poleward after landfall, as well as synoptically driven mesoscale systems.

Of additional note are the latitudinal displacements of FI maxima. Some regions, such as NGP and particularly SGP (Figs. 9b,c), have a persistent northward displacement of FI maxima relative to the forecast point; this is likely associated with the well-documented northward displacement bias of mesoscale convective systems in convection-parameterized models (e.g., Grams et al. 2006; Wang et al. 2009), including the GEFS/R, which are also responsible for many of the RPT exceedance events in these regions. In contrast, a persistent southward FI displacement is seen in the PCST and, to a lesser extent, in the SW (Fig. 9a and online supplement). This could perhaps be associated with a less-documented displacement bias of atmospheric rivers and other agents responsible for extreme precipitation in these regions (e.g., Wick et al. 2013). The FIs for the

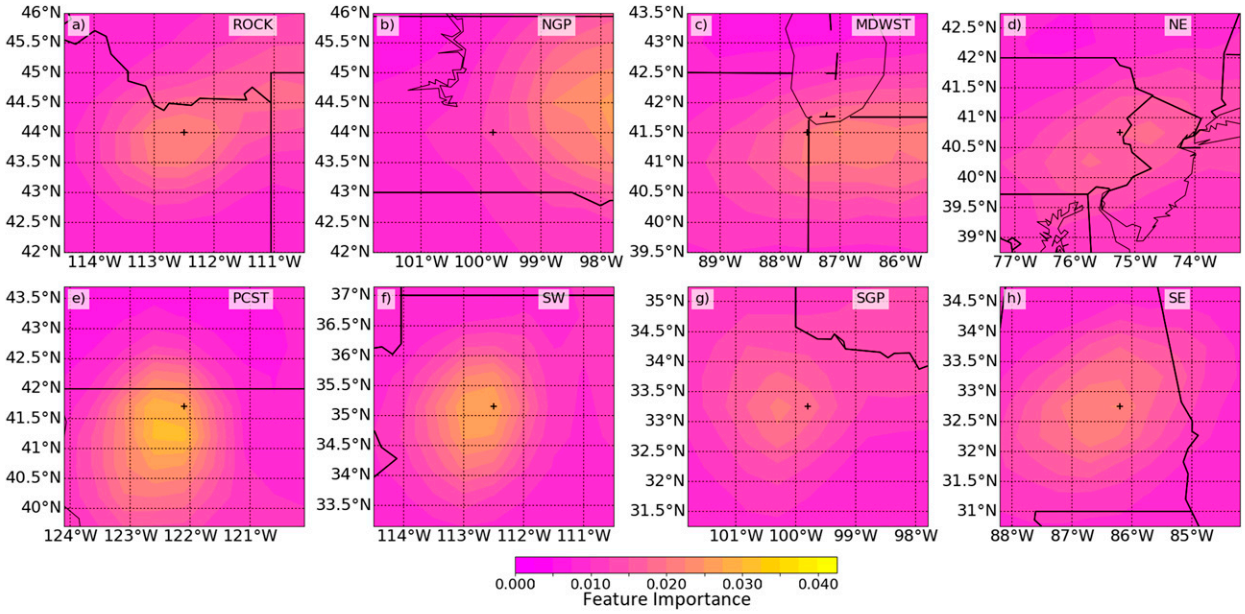

FIG. 8. Regional comparison of the summed RF FIs for the day 2 version of the CTL_NPCA model, summed over variable and time in the filled contours to give importances as a function of forecast point-relative location. Values presented correspond to the mean value obtained through fourfolds of cross validation. (a)–(h) ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions, respectively. The intersection of thick black lines indicates the location of the forecast point within each panel; other locations correspond to displaced forecast point-relative locations. Maps are drawn with the region centroid at the center of each panel, with state outlines in black underlying the panel to provide quantitative sense of spatial scale. Uniform scales are used for each panel as indicated by the figure color bars.

day 2 and day 3 versions of the model are largely quite similar (cf. Figs. 9 and 10). Many of the biases and/or displacements noted in the day 2 RF FIs remain to varying degrees in the day 3 FIs. Some differences appear to become slightly more pronounced, such as the west–east progression differences among the regions, with the PCST (Fig. 10e) shifting slightly farther west and SGP and others farther northeast, particularly at the end of the forecast period (e.g., Fig. 10g). The most pronounced difference is the general broadening of FI maxima, likely in association with increasing error and uncertainty associated with larger forecast lead times. This is suggestive of a gradual transition in trained RFs with increasing forecast lead time from bias-correcting a cohesive precipitation system simulated by the GEFS/R to predicting based on a more general characterization of the mean environment. This can be more concretely confirmed in future work by examining a wider spectrum of forecast lead times.

Interestingly and somewhat surprisingly, the PWAT FIs, shown in Fig. 11, exhibit a much different signature than the APCP FIs. In many regions, such as NGP and SGP (Figs. 11b,c), the highest PWAT FIs are located well downstream of the forecast point throughout the period. In some of these regions, such as the NE and SW (see online supplement), there is an emphasis to the east and southeast of the forecast point, whereas in others, like NGP and SGP (Figs. 11b,c), the northeast corner is favored. In some cases, the highlighted, more important portion of the domain appears to correspond to the favored moisture source for precipitation systems in the region, such as the Atlantic Ocean in the NE or the Gulf of Mexico for the SW. This is also the case for PCST (Fig. 11a), which has a persistent emphasis of importance well to the south of the forecast point; here, atmospheric rivers advect tropical moisture from the south and west. A couple of regions, in particular the SE (Fig. 11d), have a PWAT FI west–east progression like is seen for the APCP FIs in those regions. However, the PWAT FI maxima remain well to the south of the APCP FI maxima (cf. Figs. 9d and 11d), again likely capturing the source from which extreme precipitation-producing systems develop.

For the CTL_PCA model, this sort of analysis is not possible due to the transformation from feature extraction during preprocessing. However, analogous interpretation can be made through collective diagnosis of the PCs (e.g., Figs. 2–5), the relationship between the PCs and the predictand, and the FIs of the PCs themselves (Fig. 12). FI tends to decrease with increasing PC number, suggesting a correspondence between the proportion of variance in the native dataset explained by

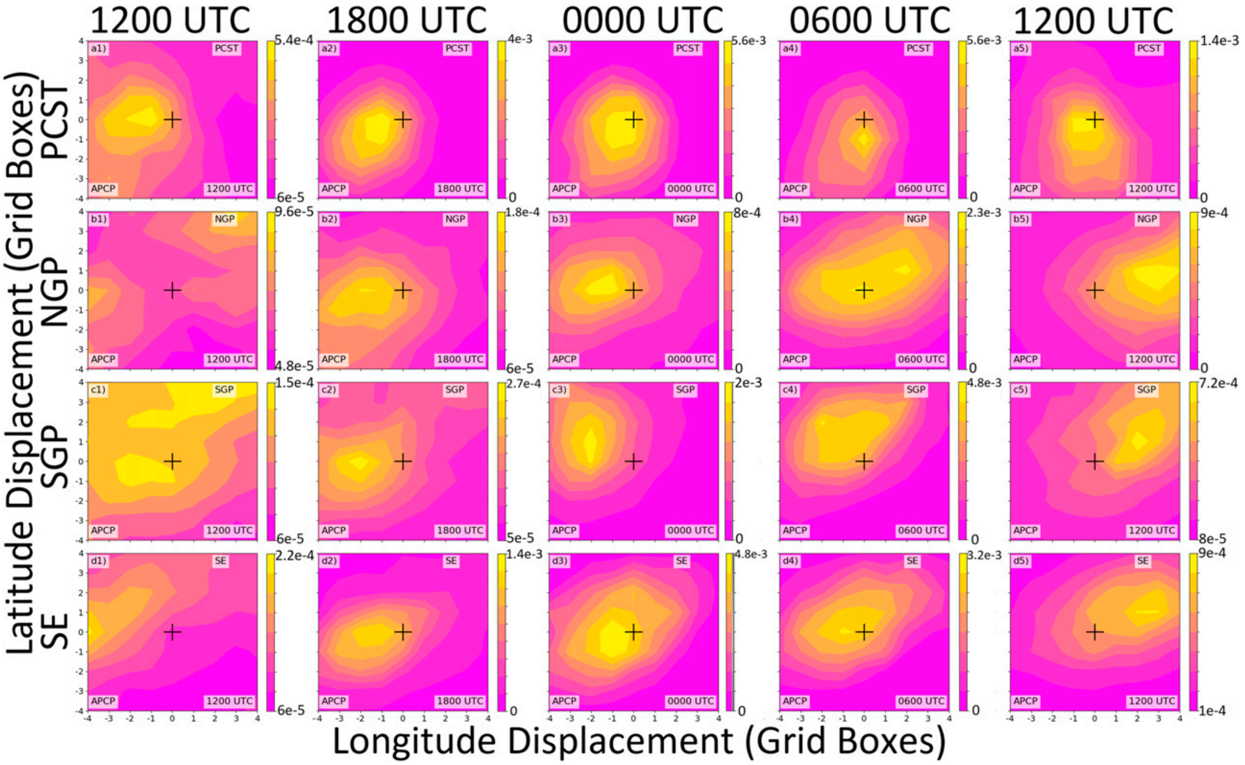

FIG. 9. Regional comparison of RF FIs for the APCP field spatially relative to the forecast point at different forecast times in the day 2 CTL_NPCA model. (a)–(d) PCST, NGP, SGP, and SE regions, respectively. (from left to right) Forecast integration hours of 36 (1200 UTC), 42 (1800 UTC), 48 (0000 UTC), 54 (0600 UTC), and 60 (1200 UTC), respectively. Values depict the mean FIs obtained through the fourfolds of cross validation. Note that the scale varies between panels; increments between colors are uniform for each color bar.

the given PC—which in turn determines its number—and the predictive ability of the PC. However, this is not uniformly the case. Every region, with the partial exception of the NGP region (Fig. 12b), has ‘‘spikes’’ in FI whereby a particular PC is identified as considerably more predictive than surrounding PCs that explain similar underlying variance. These FI maxima occur at different PC numbers depending on the region, typically somewhere between PC2 and PC15. In some regions, the first PC, which embodies the seasonal cycle (e.g., Figs. 2, 3), is by far the most predictive PC (e.g., Figs. 12b,c). On the other end of the spectrum, in PCST (Fig. 12e), the leading PC is no more predictive than much-higher-numbered PCs. In other regions still, such as ROCK, NE, and SE (Figs. 12a,d,h), PC1 is among the most predictive, but there is at least one other PC that is more predictive despite explaining less variance of the underlying forecast data. One such example is PC4 for the SE region (Fig. 12h), depicted in Fig. 13. It is associated strongly with precipitation throughout the period (Fig. 13a); anomalous moisture, especially aloft (Fig. 13g); large CIN throughout the period (Fig. 13i); and low temperature and pressure (Figs. 13b,e). It is also associated with changing surface winds, from southeasterly winds at the beginning of the period to northwesterly by the end of it, with strong spatial gradients in wind (Figs. 13c,f). As with PC2 in some regions, this again exhibits some properties consistent with frontal passage, such as drying and a switch to northerly flow advecting in from the northwest (e.g., Figs. 13f,g) and being a cool-season phenomenon (Fig. 13e), but other elements seem inconsistent, such as the lack of significant changes in temperature or pressure anomalies over the course of the period (Figs. 13b,e). With many different PCs, it can be difficult to consider all the native predictor–PC and PC–predictand relationships comprehensively, but inspection of the FIs of Fig. 12 can help target which relationships are most useful to investigate. This also allows for improved understanding of how the RF algorithm operates.

6. Results: LR diagnostics

In many cases, the CTL_LR identifies the same general findings as the RF-based models, just in a different

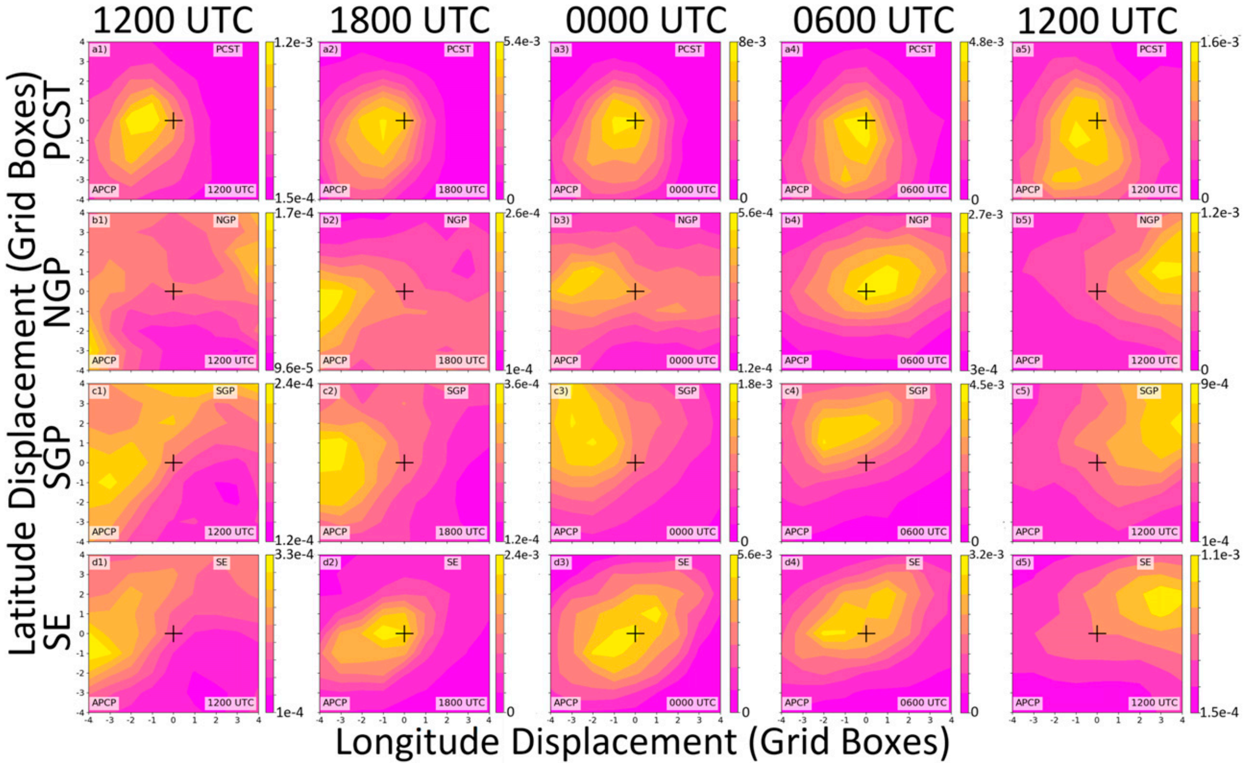

FIG. 10. As in Fig. 9, but for the day 3 model version.

capacity. One advantage of LR regression coefficients is that unlike RF FIs, they carry sign information in addition to just magnitude. Further, one can inspect coefficients for different event classes, in this case 1- versus 10-yr ARI exceedances, separately. Though there are limitations to the quantitative interpretation of the transformed regression equations, such as those for the NGP region in Fig. 14, they do still identify some important features. For the APCP field (Fig. 14a), positive coefficients unsurprisingly dominate throughout both space and time, with the one exception of the upstream side of the domain at the front-end 1200 UTC (Fig. 14a1), which actually corresponds to the 0600–1200 UTC QPF from before the start of the forecast period. But two other aspects are worthy of note. First, the coefficient maxima track the expected precipitation from the upstream to downstream side during the period, and the most positive coefficients are—like the FIs for the CTL_NPCA model—found displaced to the north of the forecast point; this is particularly evident at 0000 and 0600 UTC (Figs. 14a3,a4). Second, the coefficients are largest for the accumulations from 0000 to 1200 UTC, corresponding to the climatological peak of the diurnal cycle of extreme precipitation events in NGP (e.g., Stevenson and Schumacher 2014). Additionally, the same downstream PWAT FI maximum for the CTL_NPCA model (Fig. 11b) is reflected also in the CTL_LR model, with positive coefficient maxima downstream of the forecast point throughout the forecast period (Fig. 14d); a similar phenomenon is observed with surface moisture (Fig. 14f). It is apparent also that anomalous southeasterly flow, particularly around 0000 UTC, increases the probability of extreme precipitation events (Figs. 14g3,h3). Anomalous surface easterlies promote slower storm motion, and anomalous surface southerlies tend to yield continued moisture advection and enhanced storm maintenance (e.g., Doswell et al. 1996). Extreme precipitation event probabilities also increase with low pressure at the beginning of the period (Fig. 14i1), increasing to anomalous high pressure by the end of it (Fig. 14i5). Many extreme precipitation events in the NGP region are associated with mesoscale convective systems or other training convection. Composites of these scenarios (e.g., Peters and Schumacher 2014) have shown synoptic low pressure, particularly to the south and west of the eventual MCS, in the preconvective environment that moves out of the area or decays by the postconvective environment; this finding in the LR regression coefficients is consistent with those composites. Last, the regression coefficients somewhat counterintuitively indicate that 10-yr 24-h ARI exceedances in the NGP region are more likely with low daytime CAPE

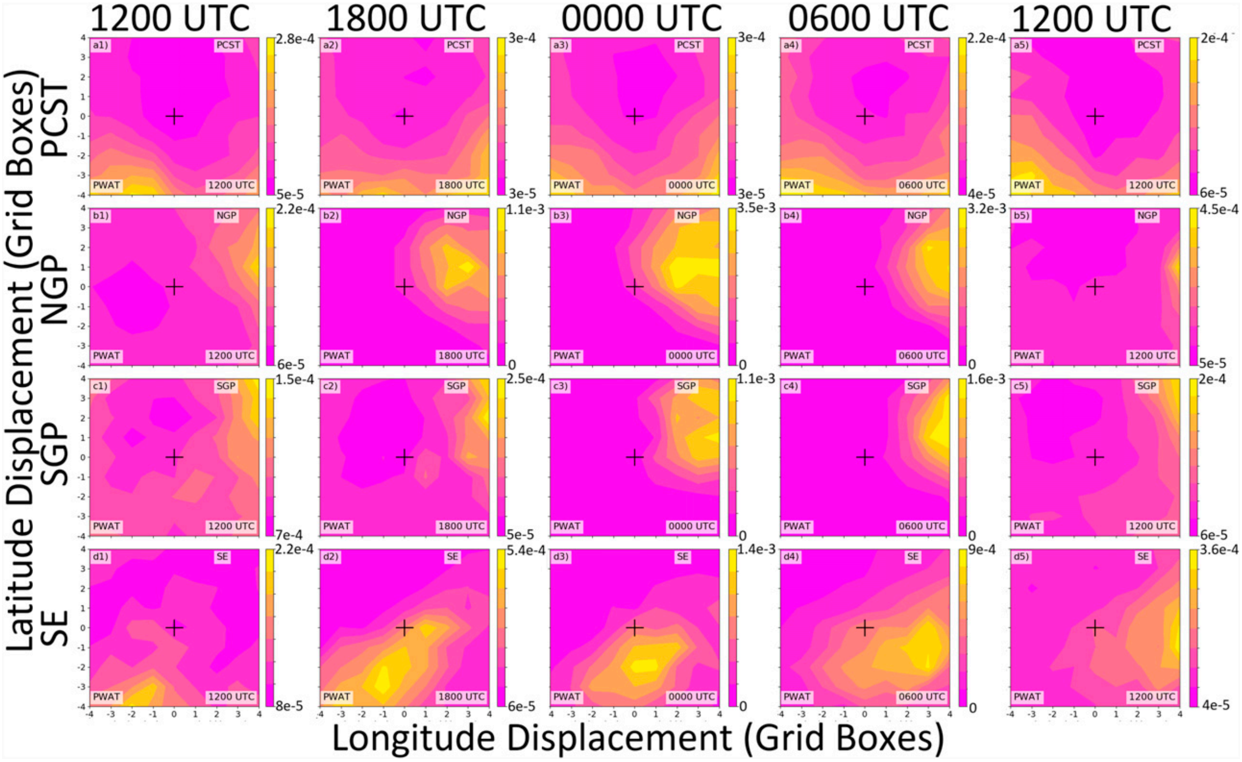

FIG. 11. As in Fig. 9, but for the PWAT field.

and high daytime CIN (Figs. 14b2,c2); this trend reverses by the end of the forecast period (Figs. 14b5,c5). This perhaps suggests that highly extreme events can occur best when instability is not exhausted from isolated diurnal convection and is instead maintained for nocturnal mesoscale convective systems that are responsible for the majority of 10-yr 24-h ARI exceedances in NGP (e.g., Schumacher and Johnson 2006).

The coefficients for the SGP region (Fig. 15) are very similar, with 10-yr exceedances associated with anomalous southeasterly surface flow (Figs. 15g3,h3), low increasing to high MSLP (Fig. 15i), and high surface and

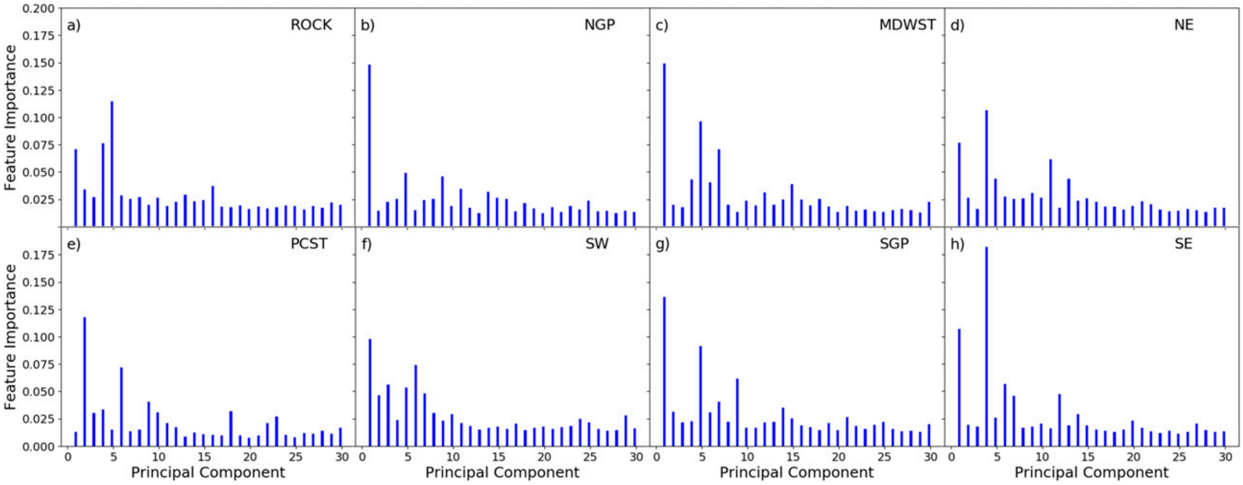

FIG. 12. Regional comparison of raw RF FIs for the day 2 version of the CTL_PCA model, shown in descending order of PC variance explained for the 30 leading PCs. Importances of background predictors exist but are omitted from this figure. (a)–(h) ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions, respectively. The scale is uniform between panels.

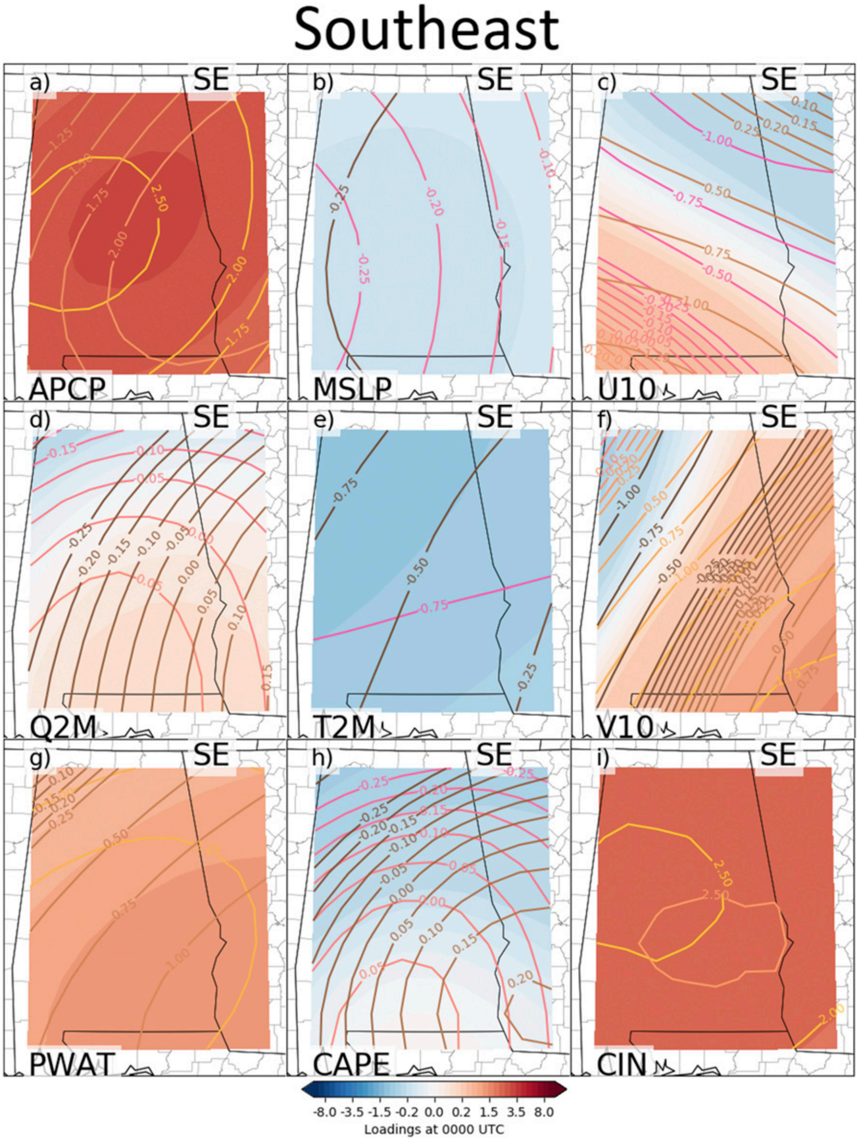

FIG. 13. As in Fig. 2, but for PC4 of the SE region.

column moisture, especially to the east and southeast of the forecast point (Figs. 15d,f). The APCP coefficients (Fig. 15a) are more spatially uniform than in NGP and have their maxima more to the south of the forecast point rather than north. The relationship with CAPE is very weak (Fig. 15b), but high CIN (Fig. 15c) to the north of the forecast point is found to correspond with SGP extreme precipitation events. These latter three variables collectively tell a similar story to NGP coefficients, but there is a redistribution of coefficient values among the fields.

Some interesting coefficient differences are observed to the east in the SE region (Fig. 16). Anomalous easterly surface flow (Fig. 16g) over the domain is again found to be conducive to extreme precipitation events; this holds to an extent with anomalous surface southerlies as well, but the coefficient values (Fig. 16h) are very small. High moisture across the forecast point domain, both throughout the surface and especially throughout the column (Figs. 16d,h) is again found to correspond to extreme precipitation events in the region. Low pressure (Fig. 16i) and temperature (Fig. 16e) tend to be positive indicators of locally extreme precipitation events. Unlike the Great Plains regions, the CAPE and CIN relationships (Figs. 16b,c) are more as expected in association with a more diurnally tied

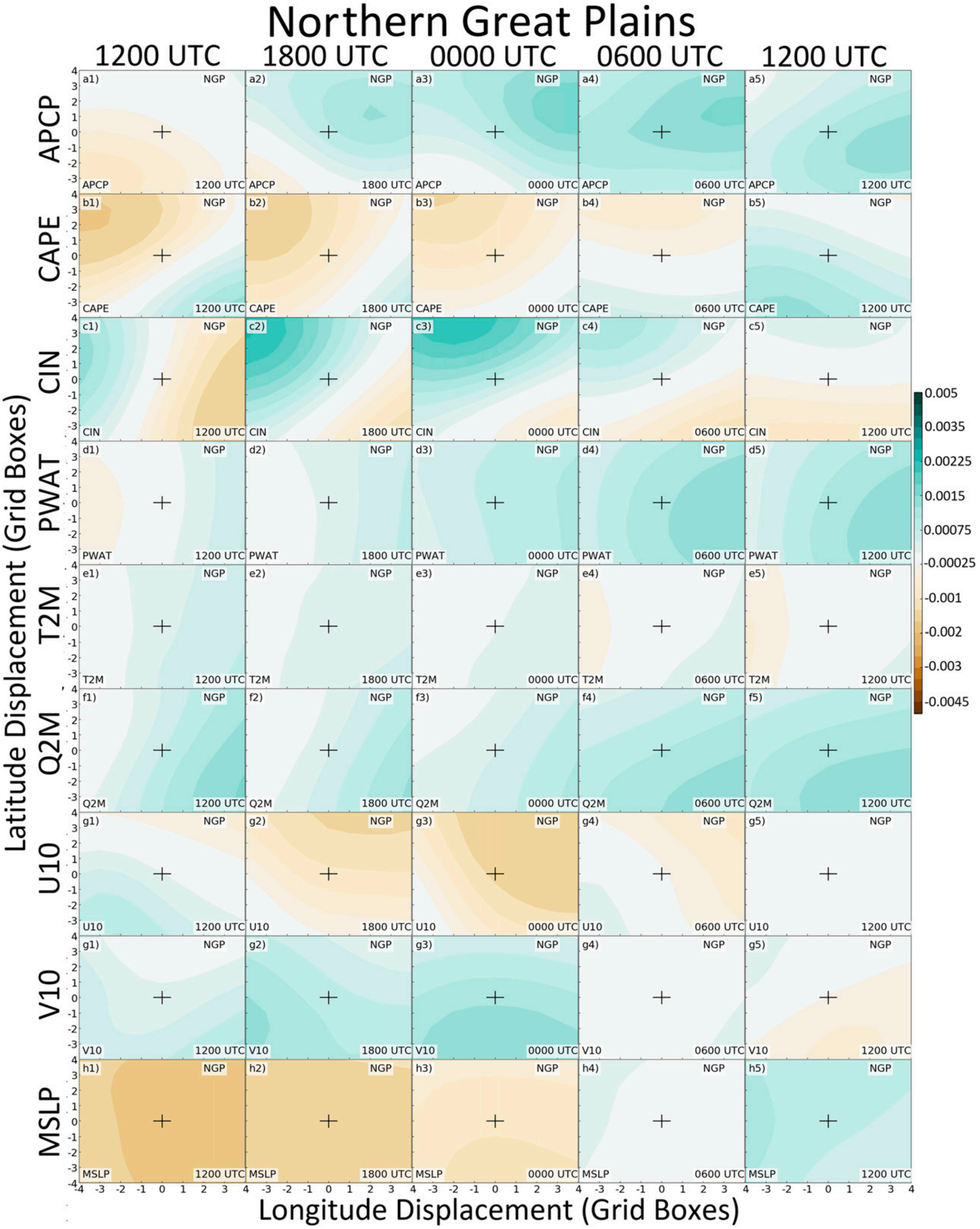

FIG. 14. Regression coefficients for the 10-yr ARI exceedance equation for the NGP region obtained through logistic regression in the day 3 version of the CTL_LR model, projected back into native variable space by means of the PC loadings. (row a)–(row i) APCP, CAPE, CIN, PWAT, T2M, Q2M, U10, V10, and MSLP forecast fields, respectively. (columns from left to right) Coefficients at 1200 UTC at the beginning of the forecast period; 1800 UTC during the period; and 0000, 0600, and 1200 UTC at the conclusion of the forecast period. Blue

precipitation climatology in the SE region, with forecasted high CAPE and low CIN during the afternoon increasing the likelihood of an extreme precipitation event. Interestingly, though high APCP corresponds with increased event probability (Fig. 16a), there is little temporal continuity of the spatial structure. What does appear to be one of the most significant indicators, as evidenced by the magnitude of the regression coefficients, is APCP to the north of the forecast point the night prior to the start of the forecast period (Fig. 16a1), which also leaves high CIN to the north of the forecast point to start the period (Fig. 16c1). This may perhaps act to favorably precondition the environment at the forecast point.

The PCST region regression coefficients (Fig. 17) yield some unusual and interesting findings that may warrant further investigation. Unlike other regions, many fields exhibit complex coefficient spatial structures, with numerous changes in sign and other smaller features. As the CTL_NPCA model identified (Fig. 6e), the CTL_LR model also identifies GEFS/R APCP as by far the most predictive field of PCST extreme precipitation events, as evidenced by the largest regression coefficients in the model occurring in Figs. 16a2 and 16a3. Also like the CTL_NPCA model, which found maximum APCP FIs to the south of the forecast point (e.g., Fig. 9a4), the same is seen in the CTL_LR coefficients (Figs. 17a2,a3). Much of the rest of the signal may be somewhat muddled because events occur most frequently in association with atmospheric river events, and these bring anomalously warm and moist conditions during the cold season. These tend to offset, leading to weaker coefficients in thermodynamic fields. But for many of these fields (e.g., Figs. 17d–f), to the extent these coefficients may be directly interpreted, low temperature and moisture at the forecast point in a surrounding environment of higher temperature and moisture tend to positively associate with extreme precipitation events in the region. This may seem rather counterintuitive, but there is some physical basis for these coefficients. In the far field, the coefficients are consistent with large-scale advection of warm, moist air over the domain, as evidenced by the increasingly positive temperature and moisture coefficients in Figs. 17d–f, and particularly PWAT (Fig. 17d). The fact that column-integrated moisture is most strongly influenced is consistent with an atmospheric river signature, where moisture is transported at mid- and upper levels and not just near the surface. But near the forecast point, where it is precipitating in the model (e.g., Fig. 17a3), there is a local minimum in temperature and moisture (Figs. 17d3,e3), consistent with column moisture condensing and precipitating out of the column; surface temperatures are likewise inhibited by a lack of radiational heating and perhaps diabatic cooling as well. Unlike the other regions, extreme events are also associated with anomalous westerly surface flow throughout the period (Fig. 17g) in this region. In other regions, easterly flow promotes slower storm motions; here, the westerly flow promotes upslope flow. Meridionally (Fig. 17h), events are associated with southerly flow transitioning to northerly flow during the forecast period, consistent with cyclone passage. Overall, some of the details of these findings may be somewhat surprising; given that, unlike most regions, the CTL_LR model had almost equal performance to the RF-based models (Herman and Schumacher 2018), this may invite deeper investigation into these properties of the coefficients.

For the interested reader, coefficients associated with the day 2 model, for unshown regions, and also for the 1-yr ARI exceedance equations have been included in the online supplement to this manuscript.

7. Summary and conclusions

Three models of different formulation from Herman and Schumacher (2018), each trained to forecast locally extreme precipitation across the CONUS, are analyzed in depth to assess their internal operations and ascertain what insights, if any, they reveal about forecasting extreme precipitation from the GEFS/R model. One model, the CTL_NPCA model, uses raw GEFS/R fields as input to a random forest algorithm to generate its predictions. The second, CTL_PCA, also uses an RF, but performs dimensionality reduction via principal component analysis on the raw GEFS/R fields and supplies a reduced predictor set consisting of just a subset of retained leading PCs in lieu of the raw fields themselves. The last, CTL_LR, also performs the PCA preprocessing step, but rather than supply the retained PCs to an RF, they are instead supplied to a regularized logistic regression algorithm. It is shown that all of these models, many of which may appear highly abstract, can be readily visualized in different ways in order to

←

values indicate the anomalously positive values of the indicated field contribute positively to the forecast probability of an ARI exceedance, while browns indicate a negative contribution. The intersection of the thick black lines indicates the location of the forecast point in each panel, with other locations depicting coefficients at spatially displaced locations.

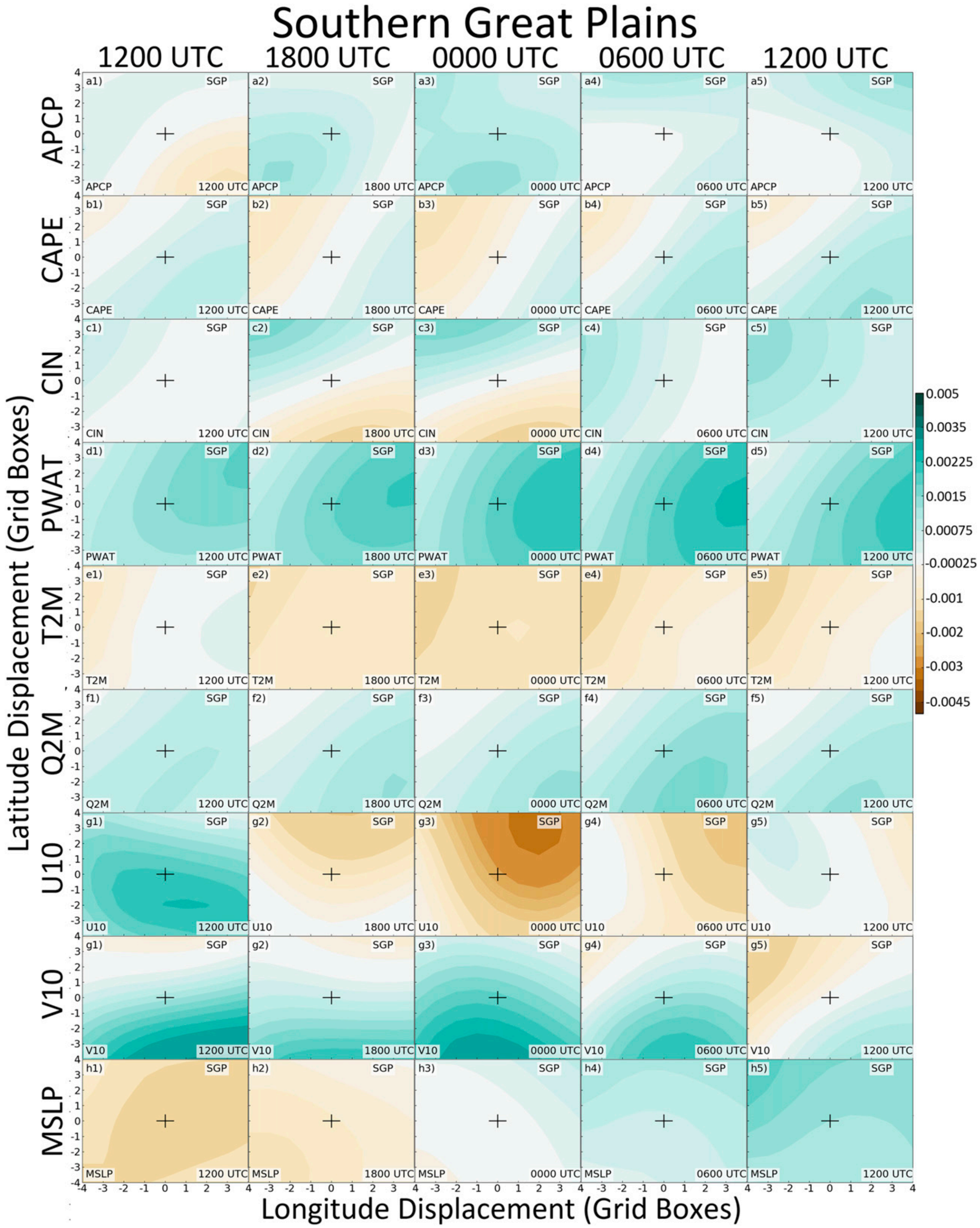

FIG. 15. As in Fig. 14, but for the SGP region.

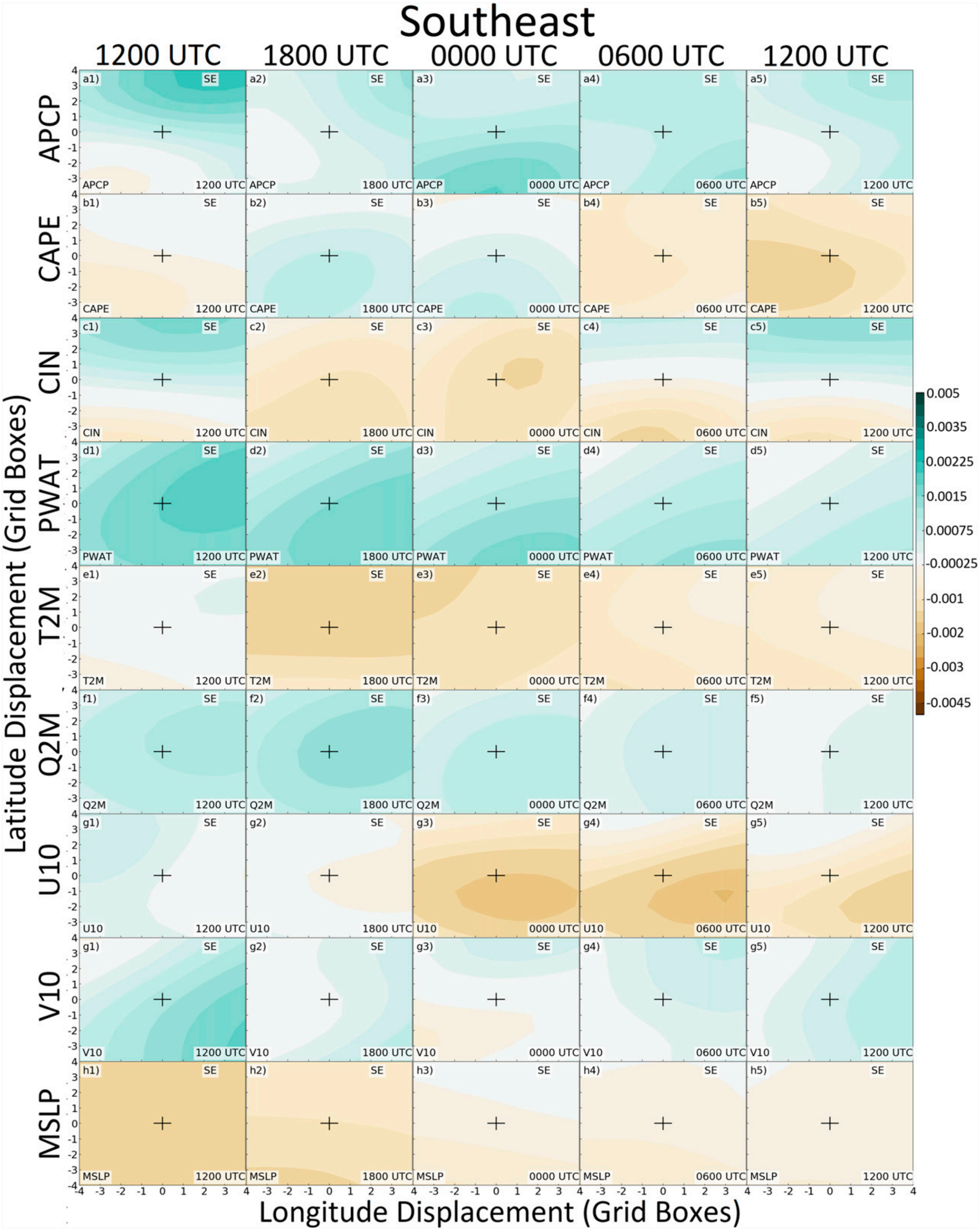

FIG. 16. As in Fig. 14, but for the SE region.

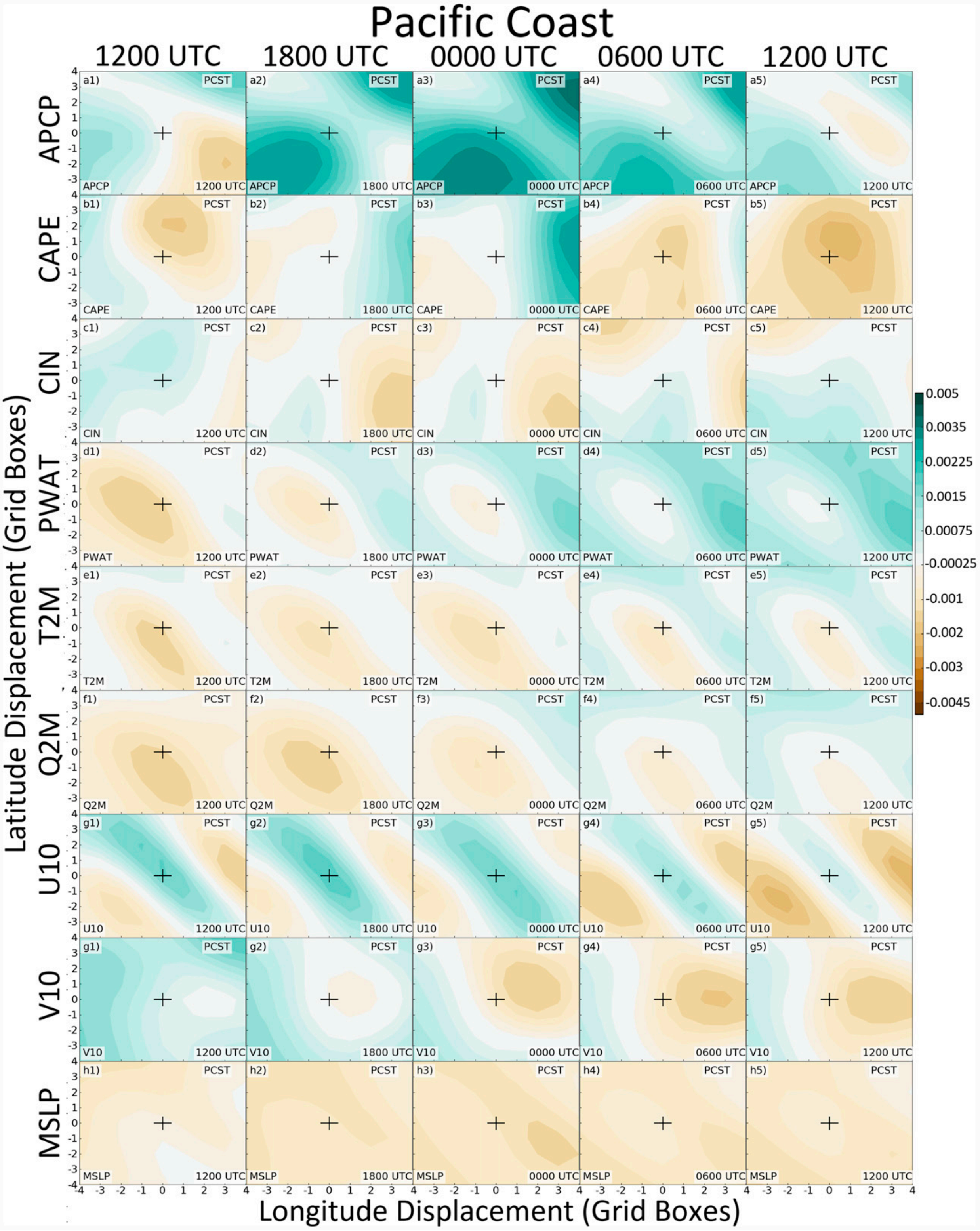

FIG. 17. As in Fig. 14, but for the PCST region.

understand their internal operations. Both the act of creating derived predictors in preprocessing via PCA and using nonparametric techniques such as RFs adds layers of abstraction that make visualization and interpretation more challenging.

Numerous aspects about forecasting locally extreme precipitation with global, convection-parameterized model output have been confirmed, while some new discoveries warrant potential further investigation. Both LR and RFs are able to identify what a human forecaster would expect to be the most predictive variables for extreme precipitation, with the largest regression coefficients and FIs generally identified for model QPFs—the direct prediction of the predictand from the GEFS/R. Moreover, the models further validate the findings of Herman and Schumacher (2016a) and other studies that found the GEFS/R and like models with parameterized convection and relatively large horizontal grid spacing have better forecasts of extreme precipitation—and in fact better QPFs all around—over the Pacific coast of the CONUS and the worst performance over the Great Plains and central CONUS. This is seen to an extent in comparing the APCP regression coefficients for the CTL_LR model, but it is especially true of the FIs in the CTL_NPCA model, which exhibit by far the highest APCP FIs in the PCST region and the lowest FIs in the NGP and MDWST regions. In fact, in the regions where extreme precipitation is most dominated by small-scale convective processes, such as NGP and MDWST (e.g., Schumacher and Johnson 2005, 2006), model QPF is not even identified as the most predictive atmospheric field from the GEFS/R, with PWAT instead exhibiting the highest FIs. Similarly, CAPE, as portrayed in the GEFS/R, is not identified to be a very predictive quantity for predicting locally excessive 24-h precipitation in most regions, but in one region, SW, where many extreme events are associated with isolated diurnally and orographically forced precipitation within monsoonal moisture, it was found to be almost equally predictive to the QPF itself. This framework and these models thus act to dynamically discern an appropriate "weighting" based on the hydrometeorology of the given region and the characteristics of the dynamical model from which the predictors are derived.

In time, the models again follow processes and focus examination dynamically depending on the region in ways consistent with how a human forecaster might approach the forecast problem. The APCP FIs follow the diurnal precipitation climatology in each region, with maxima late in the forecast period over the Great Plains and Midwest and earlier peaks over the coasts. Environmental conditions such as PWAT maximize in importance prior to the APCP FI maxima, diagnosing the relevance of these environmental properties as antecedent storm conditions. In space, the algorithm tracks precipitation features through time and space from the west edge of the predictor domain at the beginning of the period to the east edge at the end of the forecast period. Some persistent displacement biases are also noted, with a northern displacement of the maximum APCP FIs relative to the forecast point in convectively active regions such as NGP, SGP, and MDWST, in accordance with prior findings of mesoscale convective system displacement biases from convection-parameterized models (e.g., Grams et al. 2006; Wang et al. 2009; Clark et al. 2010), and a southern displacement in the PCST region, suggestive of a systematic southward displacement bias of atmospheric river events that dominate the extreme precipitation signal in the region. In aggregate, FIs are usually highest near the forecast point; however, especially in northern states east of the Rockies (NGP, MDWST, and NE), the highest mean FIs are found downstream of the forecast point. This is particularly true in the PWAT field, and the precise reasons for this identification require further investigation.

In the design of statistical forecast models, it is important to consider not necessarily just the skill of the raw model output, but the potential skill of forecasts issued by an experienced forecaster after considering the statistical model's output. If a forecast model is a complete "black box," a forecaster will inherently be unable to use knowledge of likely errors in the inputs to improve the estimate of the outcome or relate the current forecast to past scenarios where both the forecast and outcome are known, among other techniques frequently adopted by human forecasters to produce a forecast more skillful than that generated by automated guidance. With a more transparent and comprehensible model forecast process, however, a forecaster may be able to improve upon the guidance in some situations using these sorts of corrections. Of course, if a black box statistical model produces demonstrably and substantially superior forecasts to any competing guidance, it may well still outperform other less skillful models where the forecaster is able to add more value. However, as has been demonstrated in this study, machine learning algorithms including but not limited to RFs can provide forecasting insights that allow improved interpretability of the output from the statistical model, but also reveal insights about the dynamical model that allow improved interpretation of the dynamical model guidance even absent any machine learning–based model guidance. Although machine learning can identify novel properties and relationships, it should be emphasized that it is not a panacea. The diagnostics presented herein do not directly identify physical reasons for its findings; while some may be readily apparent, others require further investigation to fully understand the identified patterns in these machine learning models. That said, with existing machine learning models demonstrating considerable skill in forecasting locally extreme precipitation, as well as a host of other sensible weather phenomena, it is recommended that expected future forays into NWP with machine learning consider not only the properties of the raw forecasts that the developed models produce, but also the visualizability of the model construction and what physical insights and understanding may be gleaned from such visualization.

There are various concrete ways that these diagnostics may assist human forecasters, as well as help guide future research. Even absent using the statistical model output, these diagnostics can help a human forecaster better interpret raw dynamical guidance from the parent model—the GEFS/R, in this case. For example, the diagnostics suggest that a forecaster should shift his or her area of highest excessive precipitation risk to the south of where the heaviest precipitation is portrayed over the Great Plains and eastern regions of the CONUS, while shifting to the north along the Pacific coast. It also suggests that convective systems portrayed in the GEFS/R may be systematically too progressive, particularly in the NGP and MDWST regions—something that likely warrants further investigation. The diagnostics also help point forecasters at which fields to devote the most attention toward; in PCST, the GEFS/R's QPFs should be given considerable credence, while in NGP and MDWST, more attention should be paid to the GEFS/R PWAT field in trying to determine risk of locally extreme precipitation. The diagnostics presented in this paper provide some ability to modifying the statistical model output based on external assessments as well. For example, if a forecaster judges that the GEFS/R is much too dry aloft in a region, he or she may consult the regression coefficients and adjust probabilities accordingly, depending on the sign of the PWAT regression coefficients for the region. For RFs, if PWAT FIs are very low, the forecaster can maintain confidence in the forecast, while if they are quite high, the forecaster may choose to discount the output from the machine learning model. Additional corrections may be identifiable by performing a detailed meteorology-dependent verification of the machine learning–based forecasts over an extended historical record. A start to this type of analysis was performed in Herman and Schumacher (2018); future work should further break down model performance by meteorological regime, the analysis of which would provide even further aid to the human forecaster.

It is imperative for statistical modelers to investigate the internals of trained models to the extent possible. When performance is not appreciably degraded—and it certainly can be—it may in some instances be preferable to employ algorithms that are more easily interpretable, such as RFs in lieu of algorithms whose output is more difficult to visualize, such as support vector machines or neural networks (e.g., Rozas-Larraondo et al. 2014). Additionally, while traditional PCA was applied because the orthogonality and maximum variance constraints were believed to be beneficial for model skill and yield desirable independence properties, their potential for less physically grounded components suggests that applying more directly interpretable preprocessing instead, such as sparse PCA (Zou et al. 2006) or rotated PCA (e.g., Richman 1986; Mercer et al. 2012; Peters and Schumacher 2014), could yield more directly and easily interpretable statistical model results. Future work will seek to further explore the comparison of machine learning algorithms for NWP in additional settings as well as working to invent or apply improved methods for understanding what the machine learning informs us about the phenomenon of study.

Acknowledgments. The authors thank Joshua Hacker and two anonymous reviewers for input that helped improve the final quality of this manuscript. We also thank Tom Hamill and Gary Bates for generating and providing the GEFS/R data that made this research possible. Erik Nielsen provided helpful assistance in the creation of Fig. 1. We also wish to thank Diana Stovern, Sarah Perfater, Benjamin Albright, Mark Klein, Michael Erickson, and James Nelson at the Weather Prediction Center, who helped improve the operational utility of this research. Funding for this research was supported by NOAA Award NA16OAR4590238 and NSF Grant ACI-1450089.

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**, 581–599, https://doi.org/10.1175/WAF-D-15-0113.1.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting*, **17**, 783–799, https://doi.org/10.1175/1520-0434(2002)017<0783:COMFPQ>2.0.CO;2.

Bermingham, A., and A. Smeaton, 2011: On using Twitter to monitor political sentiment and predict election results. *Proc. Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011)*, Chiang Mai, Thailand, SAAIP, 2–10.

Bonnin, G. M., D. Todd, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2004: *Precipitation-Frequency Atlas of the United States.* NOAA Atlas 14, Vol. 1, 271 pp.

——, D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2006: *Precipitation-Frequency Atlas of the United States.* NOAA Atlas 14, Vol. 2, 301 pp.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

Cao, L.-J., and F. E. H. Tay, 2003: Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Network*, **14**, 1506–1518, https://doi.org/10.1109/TNN.2003.820556.

Clark, A. J., W. A. Gallus Jr., and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, 2011: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537.

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232, https://doi.org/10.1214/aos/1013203451.

Gagne, D. J., 2016: Coupling data science techniques and numerical weather prediction models for high-impact weather prediction. Ph.D. thesis, University of Oklahoma, 204 pp.

——, A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, ——, 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, https://doi.org/10.1175/WAF-D-17-0010.1.

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

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Grams, J. S., W. A. Gallus Jr., S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The use of a modified Ebert–McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288–306, https://doi.org/10.1175/WAF918.1.

Hall, T., H. E. Brooks, and C. A. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338–345, https://doi.org/10.1175/1520-0434(1999)014<0338:PFUANN>2.0.CO;2.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

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

——, and ——, 2016b: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea. Forecasting*, **31**, 467–482, https://doi.org/10.1175/WAF-D-15-0108.1.

——, and ——, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Hershfield, D. M., 1961: Rainfall frequency atlas of the United States. Weather Bureau, Department of Commerce Tech Paper 40, 65 pp.

Jones, T. A., D. Cecil, and M. DeMaria, 2006: Passive-microwave-enhanced Statistical Hurricane Intensity Prediction Scheme. *Wea. Forecasting*, **21**, 613–635, https://doi.org/10.1175/WAF941.1.

Larrañaga, P., and Coauthors, 2006: Machine learning in bioinformatics. *Brief. Bioinform.*, **7**, 86–112, https://doi.org/10.1093/bib/bbk007.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Statistical Forecasting Project, Department of Meteorology, MIT Science Rep. 1, 49 pp.

Marzban, C., and G. J. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626, https://doi.org/10.1175/1520-0450(1996)035<0617:ANNFTP>2.0.CO;2.

——, and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610, https://doi.org/10.1175/1520-0434(2001)016<0600:ABNNFS>2.0.CO;2.

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

Mercer, A. E., C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2012: Synoptic composites of tornadic and nontornadic outbreaks. *Mon. Wea. Rev.*, **140**, 2590–2608, https://doi.org/10.1175/MWR-D-12-00029.1.

Miller, J., R. Frederick, and R. Tracey, 1973: *Precipitation-Frequency Atlas of the Western United States*. NOAA Atlas 2, Vol. 3, 35 pp.

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, Vol. 6, 233 pp.

——, and Coauthors, 2013: *Precipitation-Frequency Atlas of the United States.* NOAA Atlas 14, Vol. 9, 163 pp.

Peters, J. M., and R. S. Schumacher, 2014: Objective categorization of heavy-rain-producing MCS synoptic types by rotated principal component analysis. *Mon. Wea. Rev.*, **142**, 1716–1737, https://doi.org/10.1175/MWR-D-13-00295.1.

Ralph, F. M., P. J. Neiman, G. N. Kiladis, K. Weickmann, and D. W. Reynolds, 2011: A multiscale observational case study of a Pacific atmospheric river exhibiting tropical–extratropical connections and a mesoscale frontal wave. *Mon. Wea. Rev.*, **139**, 1169–1189, https://doi.org/10.1175/2010MWR3596.1.

Richardson, L. F., 2007: *Weather Prediction by Numerical Process.* Cambridge University Press, 237 pp.

Richman, M. B., 1986: Rotation of principal components. *Int. J. Climatol.*, **6**, 293–335, https://doi.org/10.1002/joc.3370060305.

Roebber, P. J., 2013: Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Mon. Wea. Rev.*, **141**, 3170–3185, https://doi.org/10.1175/MWR-D-12-00285.1.

Rosten, E., and T. Drummond, 2006: Machine learning for high-speed corner detection. *Ninth European Conf. on Computer Vision*, Graz, Austria, ECCV, 430–443, https://doi.org/10.1007/11744023_34.

Rozas-Larraondo, P., I. Inza, and J. A. Lozano, 2014: A method for wind speed forecasting in airports based on nonparametric regression. *Wea. Forecasting*, **29**, 1332–1342, https://doi.org/10.1175/WAF-D-14-00006.1.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921, https://doi.org/10.1175/MWR-D-13-00168.1.

Schumacher, R. S., and R. H. Johnson, 2005: Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **133**, 961–976, https://doi.org/10.1175/MWR2899.1.

——, and ——, 2006: Characteristics of U.S. extreme rain events during 1999–2003. *Wea. Forecasting*, **21**, 69–85, https://doi.org/10.1175/WAF900.1.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, https://doi.org/10.1175/MWR3441.1.

Stevenson, S. N., and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea. Rev.*, **142**, 3147–3162, https://doi.org/10.1175/MWR-D-13-00345.1.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25, https://doi.org/10.1186/1471-2105-8-25.

——, ——, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307, https://doi.org/10.1186/1471-2105-9-307.

Thompson, D. W., and J. M. Wallace, 1998: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, **25**, 1297–1300, https://doi.org/10.1029/98GL00950.

Wang, S.-Y., T.-C. Chen, and S. E. Taylor, 2009: Evaluations of NAM forecasts on midtropospheric perturbation-induced convective storms over the U.S. northern plains. *Wea. Forecasting*, **24**, 1309–1333, https://doi.org/10.1175/2009WAF2222185.1.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Wick, G. A., P. J. Neiman, F. M. Ralph, and T. M. Hamill, 2013: Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Wea. Forecasting*, **28**, 1337–1352, https://doi.org/10.1175/WAF-D-13-00025.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, https://doi.org/10.1007/s10994-013-5346-7.

Zou, H., T. Hastie, and R. Tibshirani, 2006: Sparse principal component analysis. *J. Comput. Graph. Stat.*, **15**, 265–286, https://doi.org/10.1198/106186006X113430.