Generating Calibrated Ensembles of Physically Realistic, High-Resolution Precipitation Forecast Fields Based on GEFS Model Output

MICHAEL SCHEUERER

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

THOMAS M. HAMILL

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 5 April 2018, in final form 22 June 2018)

ABSTRACT

Enhancements of multivariate postprocessing approaches are presented that generate statistically calibrated ensembles of high-resolution precipitation forecast fields with physically realistic spatial and temporal structures based on precipitation forecasts from the Global Ensemble Forecast System (GEFS). Calibrated marginal distributions are obtained with a heteroscedastic regression approach using censored, shifted gamma distributions. To generate spatiotemporal forecast fields, a new variant of the recently proposed minimum divergence Schaake shuffle technique, which selects a set of historic dates in such a way that the associated analysis fields have marginal distributions that resemble the calibrated forecast distributions, is proposed. This variant performs univariate postprocessing at the forecast grid scale and disaggregates these coarse-scale precipitation amounts to the analysis grid by deriving a multiplicative adjustment function and using it to modify the historic analysis fields such that they match the calibrated coarse-scale precipitation forecasts. In addition, an extension of the ensemble copula coupling (ECC) technique is proposed. A mapping function is constructed that maps each raw ensemble forecast field to a high-resolution forecast field such that the resulting downscaled ensemble has the prescribed marginal distributions. A case study over an area that covers the Russian River watershed in California is presented, which shows that the forecast fields generated by the two new techniques have a physically realistic spatial structure. Quantitative verification shows that they also represent the distribution of subgrid-scale precipitation amounts better than the forecast fields generated by the standard Schaake shuffle or the ECC-Q reordering approaches.

1. Introduction

Ensemble precipitation forecasts are routinely generated at operational weather prediction centers worldwide (Molteni et al. 1996; Toth and Kalnay 1993; Charron et al. 2010) and provide valuable information about the flow-dependent forecast uncertainty. Unfortunately, systematic biases often affect all ensemble members, and not all sources of uncertainty are represented by the ensemble; it therefore cannot be considered a sample that represents the predictive distribution of observed precipitation (Park et al. 2008; Hamill et al. 2008; Bougeault et al. 2010). To obtain reliable probabilistic guidance from ensemble precipitation forecasts, a number of statistical postprocessing techniques have been proposed, including nonparametric methods such as the analog method (Hamill and Whitaker 2006; Hamill et al. 2015) or decision-tree methods (Herman and Schumacher 2018; Whan and Schmeits 2018), and parametric approaches such as extended logistic regression (Wilks 2009; Ben Bouallègue 2013), Bayesian Model Averaging (Sloughter et al. 2007; Berrocal et al. 2008; Kleiber et al. 2011), nonhomogeneous regression methods (Scheuerer 2014; Scheuerer and Hamill 2015; Stauffer et al. 2017), kernel dressing methods (Hamill and Scheuerer 2018), and methods based on a Bayesian modeling paradigm

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-18-0067.s1.

Corresponding author: Michael Scheuerer, michael.scheuerer@noaa.gov

DOI: 10.1175/JHM-D-18-0067.1

(Herr and Krzysztofowicz 2005; Wu et al. 2011; Robertson et al. 2013).

For a number of applications, the joint distribution of precipitation amounts across different locations in space and different forecast lead times is required. If, for example, the ensemble of precipitation forecasts are used as inputs to an ensemble hydrological forecast system, it is crucial that the way precipitation amounts accumulate over time and across different subbasins is represented correctly. Another type of spatial and temporal aggregation where multivariate aspects of the forecast distribution are important is the maximal precipitation amount over a certain period of time and across several forecast grid points within a certain district, a quantity that is relevant to inform severe weather warnings. One of the prevalent approaches to reconstructing the space–time variability in forecasted precipitation fields is the ensemble copula coupling (ECC) technique (Roulin and Vannitsem 2012; Schefzik et al. 2013; Flowerdew 2014), where the spatio-temporal relationships of the postprocessed precipitation fields are inherited from the raw ensemble. Another technique that is particularly common in the hydrology literature is the Schaake shuffle (Clark et al. 2004; Schaake et al. 2007), where historic observations serve as a "dependence template" for space–time variability. Several recent publications (Schefzik 2016; Scheuerer et al. 2017; Bellier et al. 2017a) have addressed one of the shortcomings of the Schaake shuffle—the lack of flow dependence of the space–time relationships—by suggesting algorithms that choose historic dates such that the weather situation at these dates is similar to the one anticipated by the forecast. In most of the articles that study multivariate probabilistic precipitation forecasts, the spatial dimension of the multivariate distribution is on the order of 10–100 (e.g., subcatchments of a river basin), and the articles considering gridded forecasts often employ standard verification metrics but do not discuss whether the postprocessed precipitation fields are physically realistic. A noteworthy exception is a recent comparison study by Wu et al. (2018), which includes example plots of the respective precipitation fields. However, the example they chose does not illustrate the inherent limitations of the different, multivariate postprocessing techniques.

In this paper we compare five different multivariate postprocessing techniques—three variants of the Schaake shuffle and two variants of the ECC technique—and assess their ability to generate ensembles of high-resolution precipitation forecast fields that are physically realistic and provide an adequate representation of the space–time variability at the analysis scale. In section 2 we describe the forecast and observation data used in this study. The statistical postprocessing methodology is explained in section 3, where we describe the approach used to generate reliable univariate forecast distributions for precipitation accumulations, review the Schaake shuffle and ECC technique, discuss their limitations with regard to generating spatially and temporally coherent precipitation forecast fields, and propose new implementations that address some of these limitations. A quantitative comparison of all four techniques is performed in section 4. We finally discuss the limitations that still exist for the two new methods and point out avenues for further improvement. All experiments and methodological development have been performed using the statistical software R (R Core Team 2017; program code is available at https://github.com/mscheuerer/PrecipitationFields).

2. Data used in this study

The postprocessing methodology discussed here is applied to forecasts of 6-h precipitation accumulations by NOAA's Global Ensemble Forecast System (GEFS) during the period from January 2002 to December 2013. Forecast data were obtained from the second-generation GEFS reforecast dataset (Hamill et al. 2013), which consists of 11 ensemble member forecasts on a Gaussian grid at ~½° grid spacing, initialized at 0000 UTC. These forecasts are calibrated and verified against climatology-calibrated precipitation analyses (Hou et al. 2014), which were obtained on a ~2.5-km grid inside the contiguous United States. We study an area in northwestern California between approximately 124.0° and 122.0°W longitude and between 38.4° and 39.8°N latitude, covering the Russian River watershed. Eleven grid points of the ½° GEFS forecast grid overlapping this area are considered, and the lower-left grid point is omitted since the associated grid cell is mainly over the ocean (see Fig. 1, first row). Likewise, we only retain the analysis grid points associated with land surfaces, which amounts to a total of 3262 grid points at the 2.5-km resolution. We consider four consecutive 6-h accumulation periods, corresponding to forecast lead times 48–54 h, 54–60 h, 60–66 h, and 66–72 h. Figure 1 depicts the forecast (first row) and analysis (third row) fields for the 24-h period that began at 0000 UTC 19 January 2010 and will be used as an example throughout the article. For the quantitative verification of the results obtained with the different postprocessing approaches, we cross-validate the 12 years of forecast and observation data, that is, we withhold one year at a time, fit the different models to the remaining 11 years' worth of data, validate the postprocessed forecast with the withheld year of independent data, and cycle through all years so that at the end 12 years' worth of out-of-sample verification cases are available. Since there is barely any

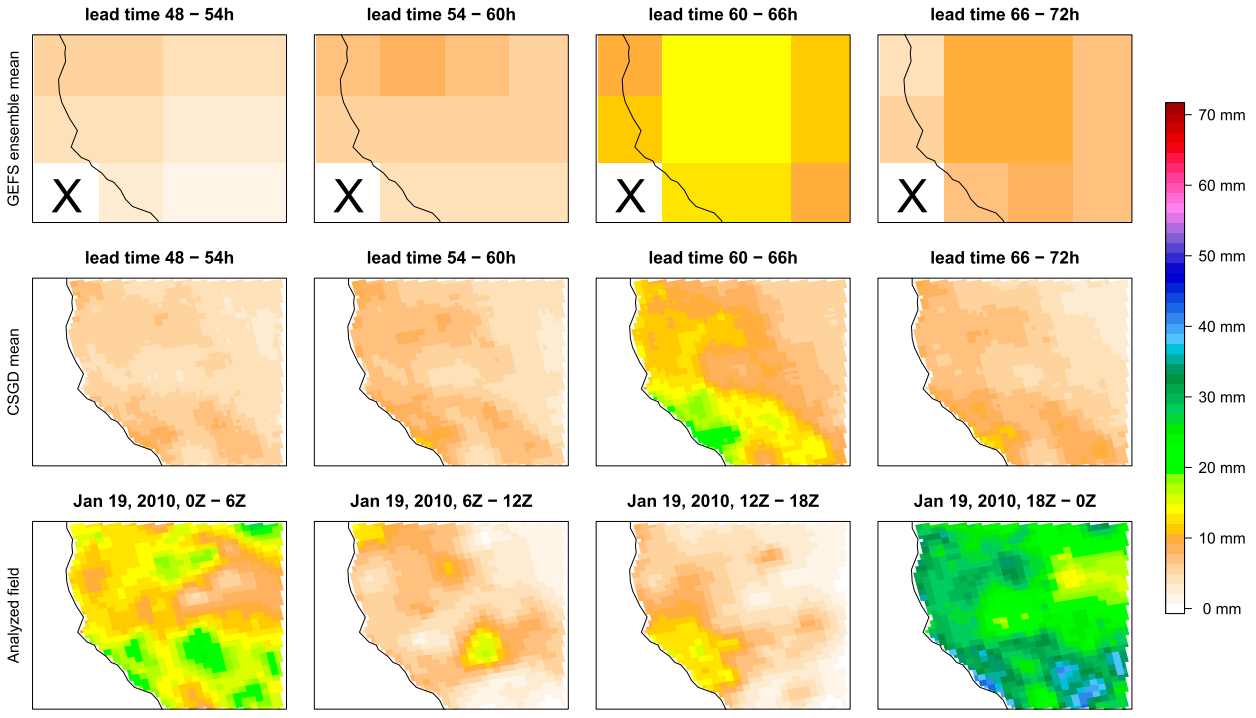

FIG. 1. (top) GEFS ensemble mean for forecast lead times 48–54 h, 54–60 h, 60–66 h, and 66–72 h, initialized at 0000 UTC 17 Jan 2010, (middle) predictive mean fields of the corresponding postprocessed forecast distributions, and (bottom) verifying analysis fields.

precipitation over this region in summer, verification is restricted to the months from October to May.

3. Statistical postprocessing methodologies

a. Univariate postprocessing

The first step in the proposed statistical postprocessing of ensemble precipitation forecasts consists of obtaining reliable, univariate forecast distributions for each analysis grid point and each lead time period. Here, we use a variant of the approach by Scheuerer and Hamill (2015) of fitting a nonhomogeneous, nonlinear regression model to training observations and statistics of the ensemble forecasts, using censored, shifted Gamma distributions (CSGDs). First, we fit a climatological CSGD to the analyzed precipitation amounts at each finescale grid point; the resulting mean, standard deviation, and shift parameters μ_{cl}, σ_{cl}, and δ_{cl} are used later in the regression equations that determine the predictive CSGD parameters. The predictors used in these equations are statistics that summarize information in the (lower resolution) raw ensemble: the ensemble probability of precipitation POP_f, the ensemble mean $\overline{f}$, and the ensemble mean absolute difference MD_f, which is a measure of ensemble dispersion. These statistics are calculated over an augmented ensemble that also comprises ensemble forecast at all forecast grid points within a certain neighborhood of the analysis grid point of interest. The size of this neighborhood and the relative weight assigned to each forecast grid point within the neighborhood is determined as described in section 3.2 of Scheuerer et al. (2017), using a data-driven weighting scheme that emphasizes forecast grid points with better predictive skill for the analysis grid point under consideration. Since the neighborhood may comprise forecast grid points with different climatologies, the associated forecasts need to be homogenized before calculating weighted means and weighted mean absolute differences. In this study, we do so by dividing each forecast by the climatological forecast mean at the respective grid point, thus replacing the ensemble forecasts by multiplicative forecast anomalies. This is much simpler but proved to be almost as effective in this application as the more complex quantile mapping procedure suggested by Scheuerer and Hamill (2015). The ensemble statistics POP_f, $\overline{f}$, and MD_f calculated from these (dimensionless) anomalies are related to the mean μ and standard deviation σ parameters of the predictive CSGD via

$$\mu = \frac{\mu_{\mathrm{cl}}}{a_1}\ \mathrm{log1p}[\mathrm{expm1}(a_1)(a_2 + a_3\mathrm{POP}_f + a_4\overline{f})]\,,$$

$$\sigma = \sigma_{\mathrm{cl}}\left(b_1\sqrt{\frac{\mu}{\mu_{\mathrm{cl}}}} + b_2\mathrm{MD}_f\right),$$

where $\log1p(x) = \log(1 + x)$, and $expm1(x) = \exp(x) - 1$. The shift parameter δ is fixed at its climatological value δ_{cl}. The regression parameters are then chosen such that the continuous ranked probability score (CRPS) obtained by applying these regression equations to a training dataset is minimized. For details about model fitting and a motivation of these equations, we refer to Scheuerer and Hamill (2015). A separate set of model parameters is fitted for each analysis grid point and each month, using data from ±45 days around the fifteenth of the respective month in the 11 years set aside for training (see section 2).

We note that for what follows, any other univariate postprocessing method that yields full predictive distributions could have been used. The results in Scheuerer and Hamill (2015) and Zhang et al. (2017), however, suggest that the CSGD approach compares favorably with other established postprocessing methods and is therefore a good choice. In the following subsections, we just assume that for each analysis grid point and each of the four 6-h-lead-time periods a calibrated, predictive distribution has been obtained by some univariate postprocessing method capable of correcting conditional biases (e.g., too many light and too few heavy precipitation events) of the raw ensemble forecasts and ensuring adequate representation of the prediction uncertainty. By calibrating the raw ensemble forecasts against the high-resolution analyzed data, the univariate postprocessing also adds spatial detail as can be seen by comparing the first two rows of Fig. 1.

b. *Schaake shuffle*

As a first reference approach to reconstructing the space–time variability of precipitation forecast fields we consider the Schaake shuffle (Clark et al. 2004; Schaake et al. 2007), which is widely used in the hydrology literature. At each grid point, this technique is applied to a sample of size K of the respective predictive distribution. To allow a direct comparison with the ECC-Q (where the "Q" stands for "quantiles") technique discussed below, which requires that K is equal to the number of raw ensemble members, we choose $K = 11$. We sample the predictive distribution systematically by taking a particular set of quantiles. This ensures that a sample of that small size represents the calibrated forecast distribution sufficiently well and has been demonstrated to improve forecast quality compared to a random sample (Schefzik et al. 2013; Wilks 2015). Specifically, we use equidistant quantiles with levels $\tau_k = (k - 0.5)/K$, $k = 1, \ldots, K$, a choice that is optimal with respect to the continuous ranked probability score (Bröcker 2012).

The Schaake shuffle is based on the idea that spatial and temporal variations in the rank-order statistics of historic analysis fields can be used to determine the space–time structure of the calibrated forecast ensemble that we aim to construct. The original approach described by Clark et al. (2004) uses the historic fields at randomly selected dates within 7 days before and after the forecast date. In the cross-validation setting used in section 4, dates can be pulled from all years between 2002 and 2013 except for the year of the forecast, and since we need $K = 11$ dates, we simply use the forecast date in those 11 years (i.e., no random selection from a larger time window). We refer to this purely date-dependent selection of historic fields as the standard Schaake shuffle (StSS) implementation. For each analysis grid point and each time period, the "shuffling" procedure now analyzes the rank order of the values of this ensemble of historic analysis fields and reorders the predictive sample in the exact same way. The resulting StSS ensemble then represents the marginal distributions obtained through univariate postprocessing and has the same rank correlations as the ensemble of historic analyses.

A major limitation of the standard implementation of the Schaake shuffle is that the historic ensemble that is used to generate rank orderings is unrelated to the anticipated weather situation. Therefore, rank correlation structures from a historic ensemble with light or moderate precipitation may be imposed on predictive samples that represent a situation with much heavier forecast precipitation. Even worse, for each analysis grid point and each forecast lead time the sample that represents the marginal distribution of a historic ensemble selected ad hoc usually contains a number $n_0 > 0$ of values with zero precipitation; these values provide no useful reordering information. If most or all values of the predictive sample are positive, the usual practice is then to order the smallest n_0 values at random, independently grid point by grid point, and independently for each lead time. The detrimental effects of this practice have been thoroughly discussed by Bellier et al. (2017a). They can also be seen in Fig. 2, which depicts the wettest member of the StSS ensemble for the 24-h period beginning at 0000 UTC 19 January 2010, and the corresponding historic analysis fields to which the values of the predictive samples were matched. Even though this is overall the wettest member, there are some areas in the historic field with zero precipitation. At grid points in those dry areas the smaller values of the predictive sample were ordered at random, independently at each finescale grid point. This produced unrealistic, noisy-looking patches in the StSS field. These patches are even larger and more frequent for the drier StSS

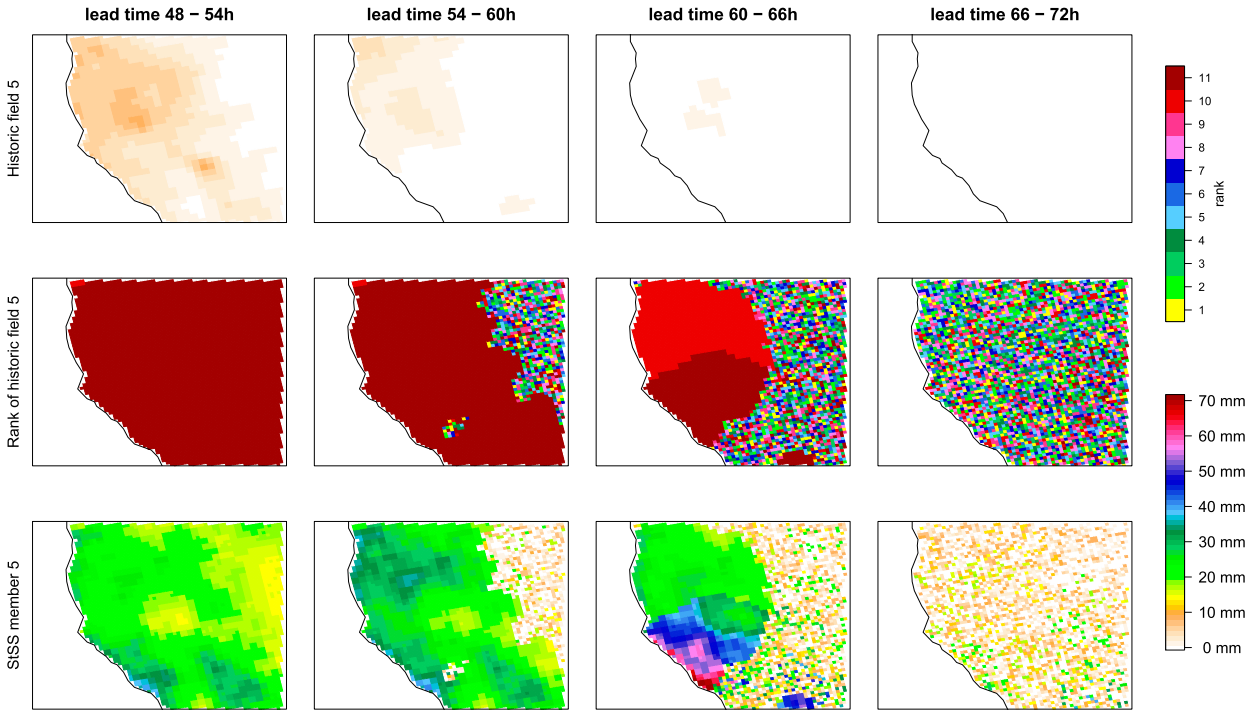

FIG. 2. (top) Historic analyzed fields corresponding to the 24-h period beginning at 0000 UTC 19 Jan 2006, (middle) ranks of the values of these historic fields among all 11 historic fields at each grid point and each lead time period, and (bottom) StSS forecast fields obtained by selecting the value of the predictive sample corresponding to this rank.

members (see online supplement A). During the third lead time period, 60–66 h, we also see an example where rank orderings in an area with light precipitation are imposed on values of the predictive sample that represent rather heavy precipitation. The values of the historic field in the southwestern part of the domain are positive but very small (the lowest precipitation category on our color scale ranges from 0 to 1.4 mm). Since almost all other historic fields had no precipitation in that area, however, these small values still assume the highest rank in the historic ensemble, and are therefore mapped to precipitation amounts of up to 70 mm. Finally, we note some discontinuities in the central part of the map as a result of the transition from the highest to the second-highest value of the respective predictive samples.

c. *Minimum divergence Schaake shuffle and quantile reordering*

To address the shortcomings of the StSS discussed above, several recent publications (Schefzik 2016; Scheuerer et al. 2017; Bellier et al. 2017a) have proposed different algorithms that make a forecast-informed selection of the historic cases used in the Schaake shuffle, with the goal of making the rank correlations inherited from the historic fields more flow dependent. The minimum-divergence Schaake shuffle (MDSS) method by Scheuerer et al. (2017) achieves this by choosing a set $\mathscr{H}$ of historic dates such that the marginal distributions of the corresponding observations closely match those of the predictive distributions across all locations (here, analysis grid points) and all forecast lead times. Specifically, denoting by F^f_{st} the cumulative distribution function (CDF) of the calibrated predictive distribution at location s and lead time t, and by $F^{\mathscr{H}}_{st}$ the empirical CDF defined by the values of the historic analysis fields corresponding to the dates in $\mathscr{H}$, we seek to minimize the sum $\Delta^{\mathscr{H}}_{\text{tot}} = \sum_{s,t} \Delta^{\mathscr{H}}_{st}$ over all divergences

$$\Delta^{\mathscr{H}}_{st} = \int \left[F^{\mathscr{H}}_{st}(x) - F^f_{st}(x)\right]^2 dx.$$

If the calibrated predictive distributions suggest that above-average precipitation amounts must be expected, the MDSS algorithm selects an ensemble of historic cases with above-average analyzed precipitation amounts. Conversely, if dry conditions are anticipated, similarity of F^f_{st} and $F^{\mathscr{H}}_{st}$ implies that $\mathscr{H}$ contains mostly historic cases with below-average precipitation amounts. In the Scheuerer et al. (2017) setup, up to seven locations and 60 lead time periods were considered, so the

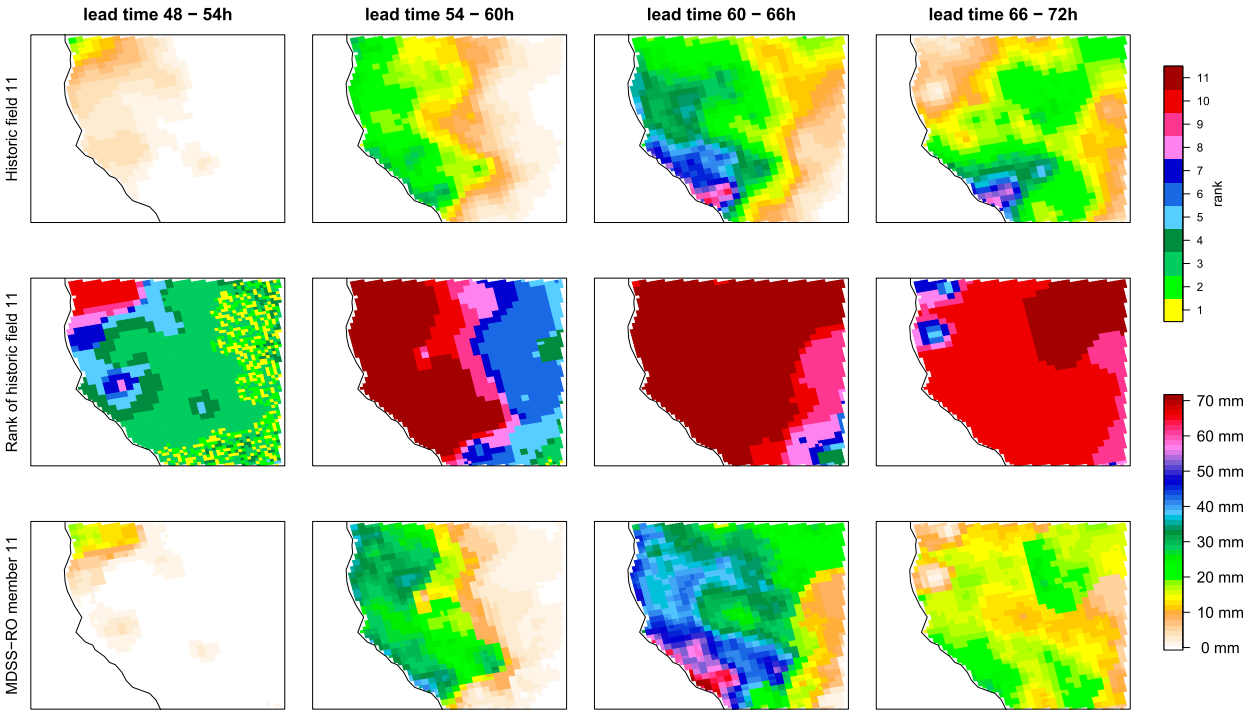

FIG. 3. (top) Historic analyzed fields associated with MDSS member 11, (middle) ranks of the values of these historic fields among all 11 historic fields at each grid point and each lead time period, and (bottom) MDSS-RO forecast fields obtained by selecting the value of the predictive sample corresponding to this rank.

dimension of the multivariate forecast distribution for precipitation was at most 420. In the present setup, we only consider four lead time periods but 3262 analysis grid points simultaneously; the resulting dimension of 13 048 is not only computationally infeasible with the original MDSS algorithm, but with only 11 years' worth of analysis data, it might also be challenging to find historic cases where the marginal distributions closely match those of the predictive distributions across all dimensions.

Therefore, in addition to performing univariate postprocessing for each of the 3262 analysis grid points, we upscale all analysis fields to the coarser resolution of the forecast grid by averaging their values over all finescale grid points associated with each forecast grid cell, and we perform univariate postprocessing also on the coarse grid. In our example, the spatial dimension thereby reduces from 3262 to 11 (the number of forecast grid cells overlapping the domain, see first row in Fig. 1). The coarse-scale predictive distributions represent a probabilistic forecast of the average precipitation amount over each forecast grid cell m and each lead time period t. The MDSS algorithm is run with these 11×4 univariate predictive distributions and upscaled historic analyses and selects K among the historic candidate dates (here we use ± 45 days around the fifteenth of the respective month $\times$ 11 training years) in such a way that the associated set of upscaled analysis fields have marginal distributions that resemble the coarse-scale predictive distributions. Once these dates have been selected, the algorithm proceeds in the same way as the standard Schaake shuffle: at each analysis grid point, the ranks of the values of the historic ensemble of analyzed fields at the selected dates are determined, and the sample from the calibrated, predictive distribution at this grid point is reordered in the same way as the historic ensemble values. We refer to this implementation as MDSS-RO because, unlike the alternative implementation suggested in the subsequent subsection, the implementation described above is purely based on sample reordering. Figure 3 shows the wettest of the 11 MDSS-RO ensemble members (see online supplement B for the full set of MDSS-RO ensemble members) for the same forecast period for which the StSS ensemble field was depicted. Since the historic cases were selected such as to be in line with anticipated precipitation amounts, the spatiotemporal structure of each member of the MDSS-RO ensemble is very similar to that of the associated historic fields; large distortions of the structure that can occur with StSS are avoided. The issue of random assignment of ranks when several values of the historic ensemble are zero is strongly reduced in wetter-than-average situations. However, it is exacerbated in dry situations, where the

values of the historic fields are actually more likely to be zero if the dates are chosen by the MDSS algorithm, as the verification in section 4 will show. Moreover, the MDSS-RO fields still have noticeable discontinuities at the locations where the ranks of the historic ensemble members change. In the following subsection, we therefore propose an alternative way of using the historic fields selected by the MDSS algorithm that addresses both the randomization and the discontinuity issue, thus yielding physically more realistic forecast fields.

d. *Minimum divergence Schaake shuffle and spatial disaggregation*

The MDSS-RO method discussed above uses the ensemble of historic analysis fields selected by the MDSS algorithm to reorder samples of the predictive distributions at the analysis grid scale. Alternatively, one can use the historic fields to spatially disaggregate forecasts of coarse-scale precipitation amounts to the finescale, and this approach is described in the following. In addition to the benefits discussed later, this MDSS-SDA (spatial disaggregation) implementation is computationally more efficient since it requires univariate postprocessing to be performed *only* at the forecast grid scale. As in section 3c, the MDSS algorithm is used to select K historic dates such that the associated set of upscaled analysis fields has marginal distributions that resemble the 11×4 coarse-scale predictive distributions. Systematic samples [quantiles with levels $(k-0.5)/K,\ k=1,\ldots,K$] of these distributions are calculated, and the reordering of the sample at each forecast grid cell and each lead time yields calibrated coarse-scale precipitation forecast fields $x^{(k)},\ k=1,\ldots,K$, with appropriate covariability between the 11 coarse-scale grid cells on the one hand and the four lead time periods on the other hand (see first two rows in Fig. 4). The next step is to disaggregate these coarse-scale fields, separately for each t, to the fine-resolution analysis grid, and we do so by using the fine-scale structure of the historic analysis fields $h^{(k)},\ k=1,\ldots,K$ selected by the MDSS algorithm (fourth row in Fig. 4). By construction, their upscaled versions have similar precipitation amounts as the reordered predictive samples. With a limited record of analyzed fields, however, we cannot expect a perfect agreement. If the values of the sample from the predictive distribution are relatively large, for example, the values of the historic ensemble are typically lower. We therefore aim to construct, for each member k, a spatially smooth, multiplicative adjustment function $\eta^{(k)}$ (third row in Fig. 4) that can be used to define an adjusted field $\tilde{h}_t^{(k)} := \eta^{(k)} h_t^{(k)}$ (fifth row in Fig. 4) such that the upscaled version of $\tilde{h}_t^{(k)}$ matches the coarse-scale field $x^{(k)}$.

Let M be the number of forecast grid cells covering the domain of interest (in our study $M=11$), and denote by υ_m the upscaling functional that maps a fine-scale field to the gridcell average corresponding to forecast grid point m. Then the requirement that the upscaled version of $\tilde{h}_t^{(k)}$ must match the coarse-scale MDSS forecast field for each forecast grid cell m can formally be written as

$$\upsilon_m(\eta^{(k)} h_t^{(k)}) = x_{mt}^{(k)}, \quad m=1,\ldots,M. \tag{1}$$

To construct a smooth function $\eta^{(k)}$ such that these M conditions are fulfilled, we make an ansatz that is common in spatial interpolation and define $\eta^{(k)}$ as a linear combination of a set of nonnegative, basis functions $\varphi_1, \ldots, \varphi_M$:

$$\eta^{(k)} = \sum_{m'=1}^{M} \alpha_{km'} \varphi_{m'}. \tag{2}$$

A visualization of our particular choice of $\varphi_1, \ldots, \varphi_M$ is given in Fig. 5, technical details are provided in appendix A. This choice is somewhat ad hoc and by no means the only option. It was guided by the following considerations:

- the basis functions should be sufficiently smooth in order for $\eta^{(k)}$ to be smooth,
- each $\varphi_{m'}$ should be compactly supported (i.e., zero outside its support radius) in order to allow adjustments that are focused on forecast grid cell m', and
- the support radius should be large enough to avoid the creation of features in $\tilde{h}_t^{(k)}$ resulting from bumps in $\eta^{(k)}$ rather than features in $h^{(k)}$.

The conditions in (1) and the representation of $\eta^{(k)}$ in (2) lead to the linear equation system

$$\sum_{m'=1}^{M} \alpha_{km'} \upsilon_m(\varphi_{m'} h_t^{(k)}) = x_{mt}^{(k)}, \quad m=1,\ldots,M, \tag{3}$$

which could, in principle, be solved for the coefficients $\alpha_{k1}, \ldots, \alpha_{kM}$. To ensure that the multiplicative adjustment function $\eta^{(k)}$ is nonnegative, however, we must require that all $\alpha_{km'}$ are nonnegative. Moreover, if coefficients associated with adjacent basis functions in Fig. 5 differ considerably, this would entail sharp gradients in $\eta^{(k)}$ and thus large distortions of the adjusted field $\tilde{h}_t^{(k)}$, so we would like adjacent coefficients to be as similar as possible. Instead of solving the equation system (3), we therefore choose $\alpha_{k1}, \ldots, \alpha_{kM}$ as the minimizers of the function

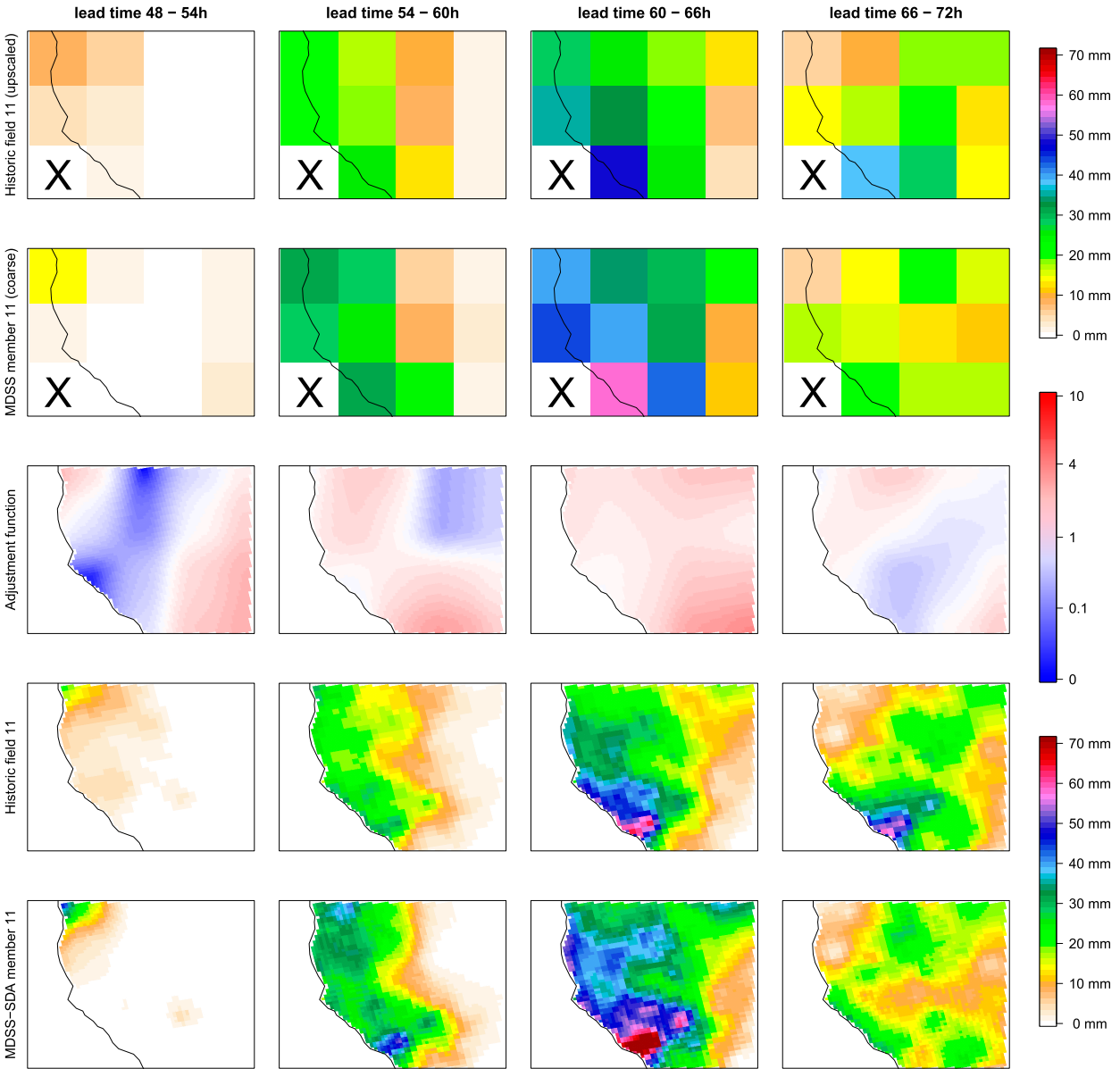

FIG. 4. Historic analyzed fields associated with MDSS member 11, (first row) upscaled and (fourth row) original; (second row) coarse-scale MDSS fields obtained by selecting, at each forecast grid point and each lead time period, the value of the coarse-scale predictive sample corresponding to the rank of the upscaled historic fields; (third row) adjustment functions with which the historic fields are multiplied at each lead time; and (fifth row) resulting high-resolution MDSS-SDA forecast field.

$$T(\alpha_{k1},\ldots,\alpha_{kM}) = \sum_{m=1}^{M}\left[\sum_{m'=1}^{M}\alpha_{km'}v_m(\varphi_{m'}h_t^{(k)}) - x_{mt}^{(k)}\right]^2 + \lambda\sum_{m=1}^{M}\left(\alpha_{km} - \frac{1}{n_m}\sum_{m'\in N_m}\alpha_{km'}\right)^2,$$

where N_m is the set of forecast grid points adjacent to grid point m, and n_m is the size of this set. The first term of the minimization target T is zero if the equation system (3) is solved exactly and increases with increasing difference between the upscaled version of $\tilde{h}_t^{(k)}$ and $x^{(k)}$. The second term is the sum of squared differences between each α_{km} and the average over its neighbors, and therefore penalizes volatility of $\eta^{(k)}$. The degree of penalization is controlled by the parameter λ, which in this study was chosen to be $\lambda = 0.25$. In our setup, this value provides a good trade-off between the smoothness of $\eta^{(k)}$ and the quality of the representation of the calibrated predictive distributions, which degrades as the first term of T is de-emphasized. The judgement of

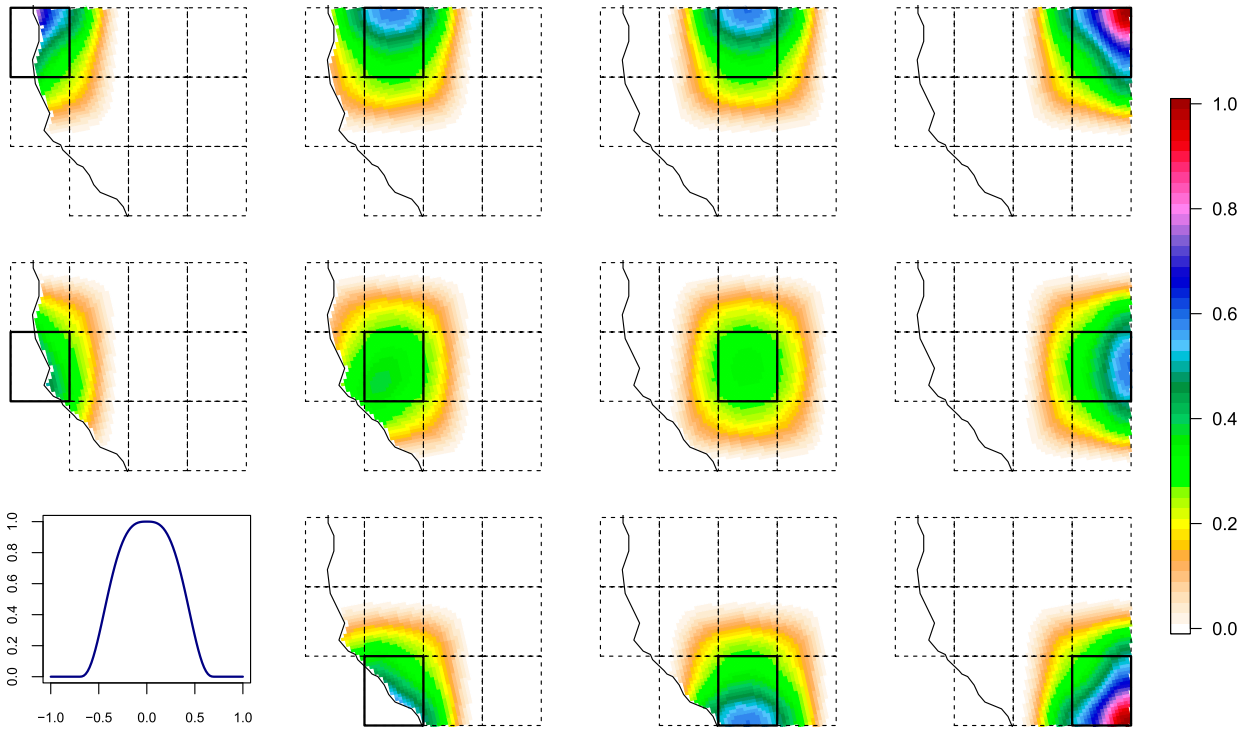

FIG. 5. Basis functions $\varphi_1, \ldots, \varphi_{11}$ used in (2). The plot in the lower-left corner depicts the tricube kernel used to construct these basis functions (see appendix A for details).

whether $\eta^{(k)}$ is sufficiently smooth and the resulting adjusted field is realistic is rather subjective. It is therefore difficult to determine λ in an objective way unless a verification metric can be found which quantifies the physical consistency of the postprocessed precipitation forecast fields. Minimization of T is performed under the constraint that all coefficients are nonnegative, which can be done, for example, by using the limited-memory Broyden–Fletcher–Goldfarb–Shanno for bound constraints (L-BFGS-B) algorithm (Byrd et al. 1995) implemented in R. Calculating $\alpha_{k1}, \ldots, \alpha_{kM}$ in this way does not guarantee that the upscaled version of $\tilde{h}_t^{(k)}$ matches $x^{(k)}$ exactly, but we usually get a very good approximation of it while enforcing a reasonably smooth adjustment function $\eta^{(k)}$ and thereby an adjusted field $\tilde{h}_t^{(k)}$ that is physically more realistic. The full set of MDSS-SDA ensemble members constructed in the setup of our case study is provided in online supplement C.

e. *Ensemble copula coupling*

All of the Schaake shuffle variants discussed above use an ensemble of historic analysis fields to reconstruct the space–time variability of postprocessed forecast fields. Alternatively, their spatiotemporal structure can be modeled after the raw ensemble forecasts (Roulin and Vannitsem 2012; Schefzik et al. 2013; Flowerdew 2014), and this approach has been referred to as ensemble copula coupling. In the present case where the grid spacing of the raw ensemble is much coarser than that of the analysis against which it is calibrated, subgrid-scale variability in in the spatiotemporal structure will not be represented as well as in an analysis-based modeling approach (if this small-scale variability in the analysis is trustworthy). Space–time variability at the forecast grid scale, however, should be represented well since the raw ensemble is based on a physical model.

As with the StSS approach, we start by generating a full predictive distribution at each analysis grid point and each forecast lead time, and sample this distribution by calculating the quantiles with levels $\tau_k = (k - 0.5)/K$, $k = 1, \ldots, K$. Schefzik et al. (2013) refer to this particular implementation as ECC-Q, and note that it has better sampling properties than ECC-R, which is based on a random sample. The sample size K must be equal to the number of raw ensemble members, which in the present setup means $K = 11$. The next step of ECC-Q is again similar to the StSS method: at each analysis grid point and each lead time the predictive sample is reordered in the same way as the raw ensemble values. For this purpose, the raw ensemble forecast fields are bilinearly interpolated to the analysis grid. Akin to StSS, it is possible that a number $n_0 > 0$ of the raw ensemble forecast values are zero, and thus do not provide any information on the rank order. Resolving these ties at random can result in

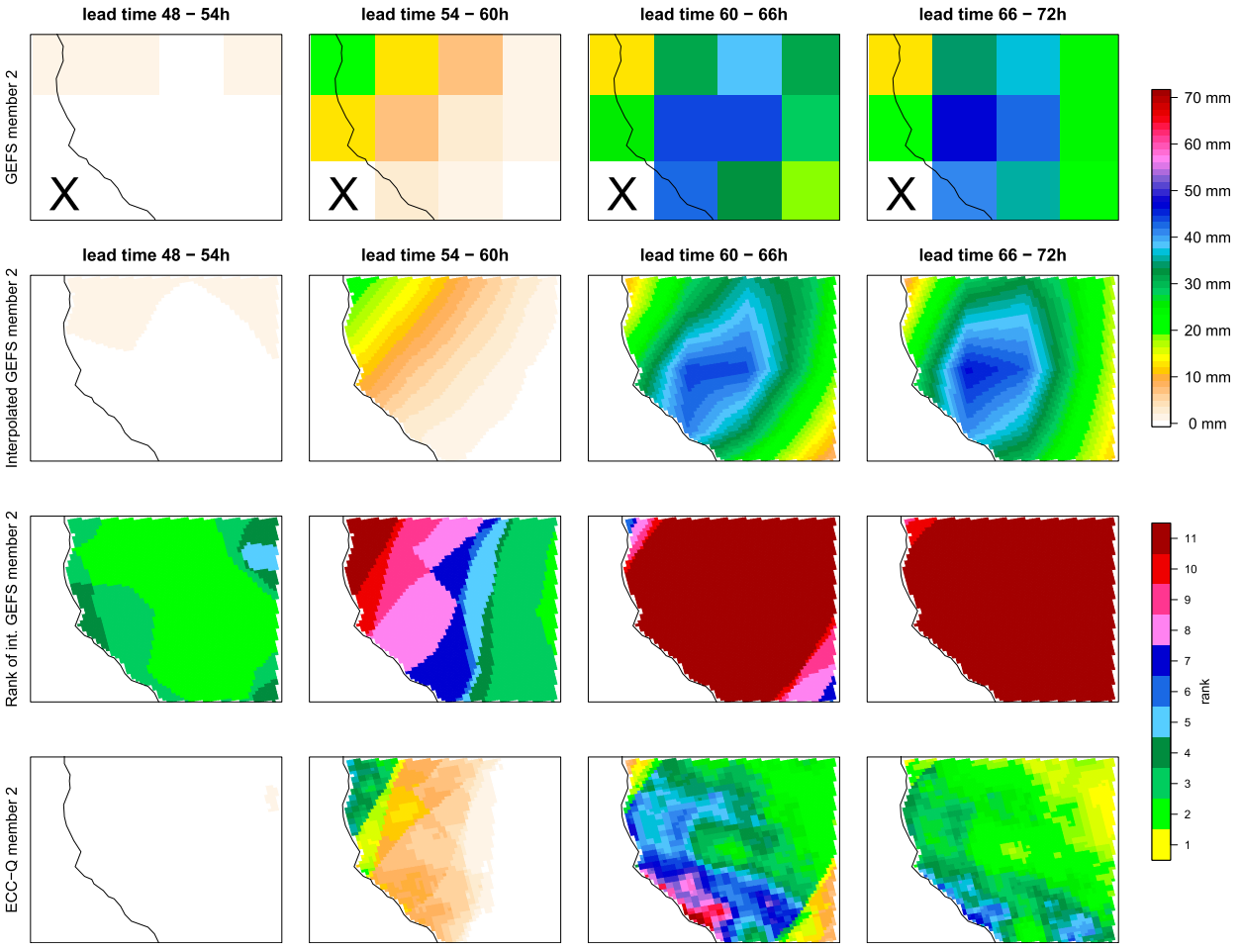

FIG. 6. (first row) Original and (second row) interpolated GEFS member 2 of the ensemble forecast initialized at 0000 UTC 17 Jan 2010, (third row) associated rank at each analysis grid point and each lead time period, and (fourth row) ECC-Q forecast fields obtained by selecting the value of the predictive sample corresponding to this rank.

noisy-looking patches in the ECC-Q forecast fields similar to those noted above in Fig. 2. In contrast to StSS, however, the conditions under which this tends to happen are reversed. For StSS, the number n_0 of ensemble members for which no ordering information is available is approximately K times the climatological probability of zero precipitation, since the historic analysis fields are selected unconditionally. In moist forecast situations like the one in our case study, the number of nonzero values in the predictive sample is typically larger than n_0, and this is therefore the situation where random reordering occurs most often. For ECC-Q, on the contrary, a large number of nonzero values in the (calibrated) predictive sample usually goes along with a large number of nonzero raw ensemble members, and so random reordering rarely occurs in moist forecast situations. It is common though in dry forecast situations, when the raw ensemble is overconfident and all or almost all members forecasts are zero, but one or two values of the predictive sample are nonzero. This case does not occur for our example forecast date, but we will see some effects of this random reordering in dry situations in section 4.

Figure 6 shows the interpolated forecast fields corresponding to GEFS member 2 (which resulted in the wettest ECC-Q ensemble member), the associated ranks, and the resulting ECC-Q forecast fields. The interpolated GEFS forecasts (first row) do not provide any information on subgrid-scale variability; the subgrid-scale features in the plots result from the univariate postprocessing, which implicitly accounts for climatological differences at the subgrid scale. Figure 6 reveals another flaw in the ECC-Q forecast fields (third row, especially for lead time 54–60 h): a number of discontinuities can occur where two of the interpolated raw ensemble members intersect and change their rank order within the ensemble. While the predictive distributions—and thus their quantiles—are continuous over the domain, the integer-valued rank fields (third row in

Fig. 6) are discontinuous. Any change of ranks entails jumping from a lower to a higher (or vice versa) quantile, and this is what causes the physically unrealistic appearance of some of the ECC-Q forecast fields (see online supplement D for all 11 ECC-Q ensemble members). If the ensemble size was larger, the differences between the values of the predictive samples would be smaller and the jumps would be less apparent, but they would never disappear entirely since the quantile fields never intersect, and so changing from one quantile field to another inevitably entails a discontinuity of the ECC-Q ensemble fields.

Schefzik et al. (2013) discuss a third variant, ECC-T (where the "T" stands for "transformations"), in addition to ECC-Q and ECC-R. Instead of reordering quantiles or random samples, ECC-T uses a quantile-mapping procedure that constructs—separately for each analysis grid point and each lead time—a function that maps the values of the raw ensemble to values that represent the calibrated predictive distribution. This is achieved by fitting a parametric distribution to the (interpolated) raw ensemble forecasts and using the inverse CDF of this distribution to define the quantile levels at which the predictive distribution is evaluated, in contrast to ECC-Q, where fixed quantile levels are selected at each grid point. The disadvantage of this procedure is that the ECC-T quantile levels typically result in a less optimal representation of the predictive distribution. Its major benefit is that similar values of the raw ensemble are mapped to similar values of the calibrated ensemble, and this would potentially avoid the discontinuities seen in Fig. 6. The particular implementation proposed by Schefzik et al. (2013), however, poses significant challenges in the setup studied here, since a sufficiently flexible, parametric distribution would have to be fitted to the forecast values of only 11 raw ensemble members, and in a lot of cases most or all of these values are zero. Under these circumstances, an adequate fit is not guaranteed. A recent paper by Bellier et al. (2018) compares a variant of ECC-T with alternative approaches like trajectory smoothing in the context of multivariate hydrological postprocessing where similar challenges (ties in the raw ensemble forecasts, discontinuities) arise and concludes that the ECC-T trajectories are not competitive, even though the ensemble size in their setup is much larger ($K = 51$). In the subsequent section, we propose a variant of ECC-Q that makes two modifications and combines the idea of reordering a quantile-based sample with the ECC-T idea of constructing a mapping function that maps the values of the interpolated raw ensemble forecasts to a sample of calibrated forecasts. Unlike ECC-T, however, our approach is capable of dealing with the challenges that come with a small ensemble size and the complex nature of the distribution of ensemble forecasts for precipitation.

f. Modifications to ECC-Q

We propose two modifications to the ECC-Q technique that address the shortcomings mentioned above. The first one addresses the issue of random reordering of predictive samples when several members of the raw ensemble predict zero precipitation by simulating "negative precipitation." For each member k, let S_k be the set of forecast grid points where this member predicts zero precipitation. To permit bilinear interpolation of the modified ensembles, we consider forecast grid points over an area that extends slightly beyond the boundaries of the domain of study, so that the number of grid points considered is now $\tilde{M} > M$. Spatial correlations between different forecast grid points are modeled through basis functions $\tilde{\varphi}_m$, which are constructed in the same way as those described in section 3d and appendix A, except that we now consider the extended set of forecast grid points and choose the support radius ρ to be 3 times the forecast grid spacing, that is, ~1.5°. This means that the spatial correlations of our simulated negative precipitation are zero beyond a distance of approximately 150 km. This choice is ad hoc and could certainly be optimized or made more weather dependent, but such improvements are not straightforward and are beyond the scope of this paper. The simulation now proceeds as follows:

1) define coefficients $\varepsilon_1, \ldots, \varepsilon_{\tilde{M}}$ by simulating ε_m from a uniform distribution $\mathscr{U}_{[-1,0]}$ for all $m \in S_k$ and setting $\varepsilon_m = 0$ otherwise;
2) define the negative precipitation field $\tilde{x}^{(k)} := \sum_{m=1}^{\tilde{M}} \varepsilon_m \tilde{\varphi}_m$;
3) at each forecast grid point where the value of the raw ensemble member k is zero, replace it by the value of $\tilde{x}^{(k)}$ at this grid point; and
4) bilinearly interpolate the resulting modified raw ensemble forecasts to the analysis grid.

The resulting modified interpolated raw ensemble members now have spatially correlated, negative values instead of zeros and thereby avoid the randomization entailed by tied ranks in the ECC-Q reordering procedure. The spatial correlations implied by the basis functions $\tilde{\varphi}_1, \ldots, \tilde{\varphi}_M$ are not based on any statistical or physical model, but the results in section 4 show that they are more appropriate than the spatial independence assumption entailed by randomization of tied ranks.

Consider now a fixed analysis grid point and lead time at which the ECC-Q procedure is applied. Denote by $z_1, \ldots, z_K$ the values of the interpolated raw ensemble

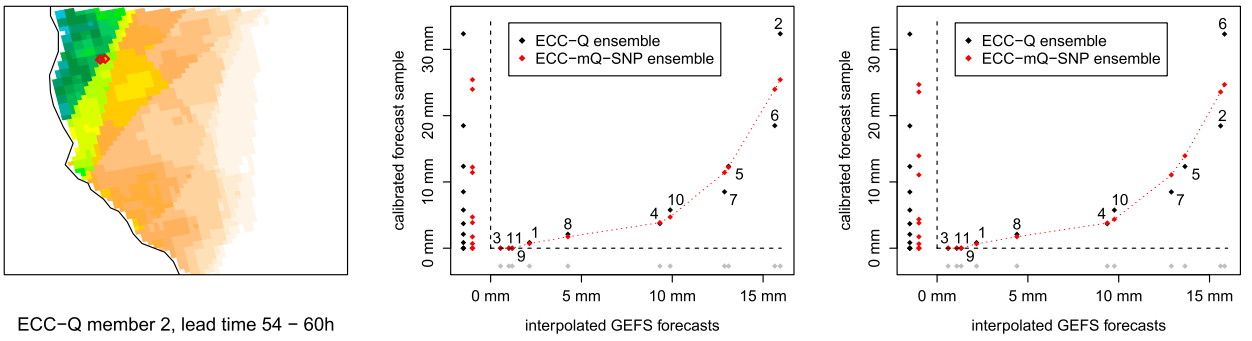

FIG. 7. (left) Example of an ECC-Q forecast field with discontinuities and (center),(right) illustration of the ECC-mQ-SNP mapping function (red curves) at two adjacent analysis grid points (red diamonds in the left plot). The numbers in black denote the respective ensemble member.

forecasts and by $\tilde{z}_1, \ldots, \tilde{z}_K$ the ECC-Q ensemble obtained by reordering the sample [quantiles with levels $\tau_k = (k - 0.5)/K,\ k = 1, \ldots, K$] that represents the calibrated predictive distribution. If we look at the ECC-Q ensemble from an ECC-T perspective, we can consider the reordering the result of a mapping $\tilde{\psi}: z_k \mapsto \tilde{z}_k$ for $k = 1, \ldots, K$. Unlike the ECC-T quantile mapping procedure which implies that two interpolated raw ensemble members with similar values are mapped to two ECC-Q ensemble members which have again similar values, the ECC-Q mapping function $\tilde{\psi}$ is discontinuous as illustrated in Fig. 7 for two analysis grid points in an area where the values of two interpolated raw ensemble fields (2 and 6) intersect. Since these grid points are adjacent, the values of predictive sample (black diamonds) are very similar. The values of the interpolated raw ensemble fields at these two grid points are also very similar, but at one of them the value of member 2 is slightly larger and at the other one the value of member 6 is slightly larger. As a result, the ECC-Q member 2 forecast field abruptly jumps from the highest to the second-highest sample value and thereby creates a discontinuity. To fix this, we propose a way of constructing a modified mapping function $\hat{\psi}: z_k \mapsto \hat{z}_k,\ k = 1, \ldots, K$ such that

- for each k, $\hat{\psi}(z_k) \approx \tilde{z}_k$ as far as possible, while
- $\hat{\psi}$ does not have steep gradients, that is, if $z_k \approx z_{k'}$ then $\hat{z}_k \approx \hat{z}_{k'}$.

The first point is desirable because $\tilde{z}_1, \ldots, \tilde{z}_K$ is a CRPS-optimal sample, and we would like to preserve this property as much as possible. On the other hand, we have to modify the sample values somewhat to achieve continuity of $\hat{\psi}$ as a mapping function because this in turn warrants spatial continuity of the calibrated ensemble forecast fields. This is achieved by defining $\hat{\psi}$ as a linear regression spline that is fitted to the ECC-Q point pairs $(z_1, \tilde{z}_1), \ldots, (z_K, \tilde{z}_K)$ while discouraging steep gradients of this spline. If the values $\tilde{z}_1, \ldots, \tilde{z}_K$ are all zero, we can simply set $\hat{\psi} \equiv 0$. If only one value $\tilde{z}_k$ is nonzero, we set $\hat{\psi}(z_k) = \tilde{z}_k$ and $\hat{\psi}(z_j) = 0$ for all $j \neq k$. Otherwise, let π be the permutation of indices such that $z_{\pi(1)} < \cdots < z_{\pi(K)}$ and let $\tilde{n}_0$ be the number of zeros in $\tilde{z}_1, \ldots, \tilde{z}_K$. If all $\tilde{z}_k$ are nonzero, we set $\tilde{n}_0 = 1$ (this is necessary for the spline to be well defined). We now define $\hat{\psi}$ as a linear regression spline with knots $z_{\pi(n_0+1)}, \ldots, z_{\pi(K-1)}$:

$$\hat{\psi}\left(z; \gamma_0, \ldots, \gamma_{K-\tilde{n}_0}\right) = \gamma_0 + \gamma_1 z + \sum_{j=\tilde{n}_0+1}^{K-1} \gamma_{j-\tilde{n}_0+1}\left(z - z_{\pi(j)}\right)_+, \tag{4}$$

where $(\cdot)_+ = \max(\cdot, 0)$. The first two terms of $\hat{\psi}$ define a standard linear regression function while the additional terms allow the slope of the fitted curve to increase or decrease at every knot. The regression coefficients $\gamma_0, \ldots, \gamma_{K-\tilde{n}_0}$ are determined by minimizing

$$\hat{T}\left(\gamma_0, \ldots, \gamma_{K-\tilde{n}_0}\right) = \sum_{k=\tilde{n}_0}^{K}\left[\hat{\psi}\left(z_{\pi(k)}; \gamma_0, \ldots, \gamma_{K-\tilde{n}_0}\right) - \tilde{z}_{\pi(k)}\right]^2 + \hat{\lambda}\sum_{k=\tilde{n}_0+1}^{K-1} \max(\tilde{z}_{\pi(k)}, 1)\gamma_{k-\tilde{n}_0+1}^2. \tag{5}$$

The first sum in (5) measures the accuracy of the approximation $\hat{\psi}(z_k) \approx \tilde{z}_k$. If $\hat{\lambda} = 0$, coefficients can be found such that the approximation is exact, and the resulting mapping function $\hat{\psi}$ would entail the same mapping as the ECC-Q approach. For $\hat{\lambda} > 0$, the second sum in (5) penalizes large values of the coefficients $\gamma_2, \ldots, \gamma_{K-\tilde{n}_0}$, and thereby discourages large changes in the slope of $\hat{\psi}$. The penalty for coefficient $\gamma_{k-\tilde{n}_0+1}$ increases as $\tilde{z}_{\pi(k)}$ increases to account for the fact that larger values of $\tilde{z}_{\pi(k)}$ typically go along with larger approximation errors $\psi(z_{\pi(k)}; \gamma_0, \ldots, \gamma_{K-\tilde{n}_0}) - \tilde{z}_{\pi(k)}$, and therefore a larger penalty is required to balance that.

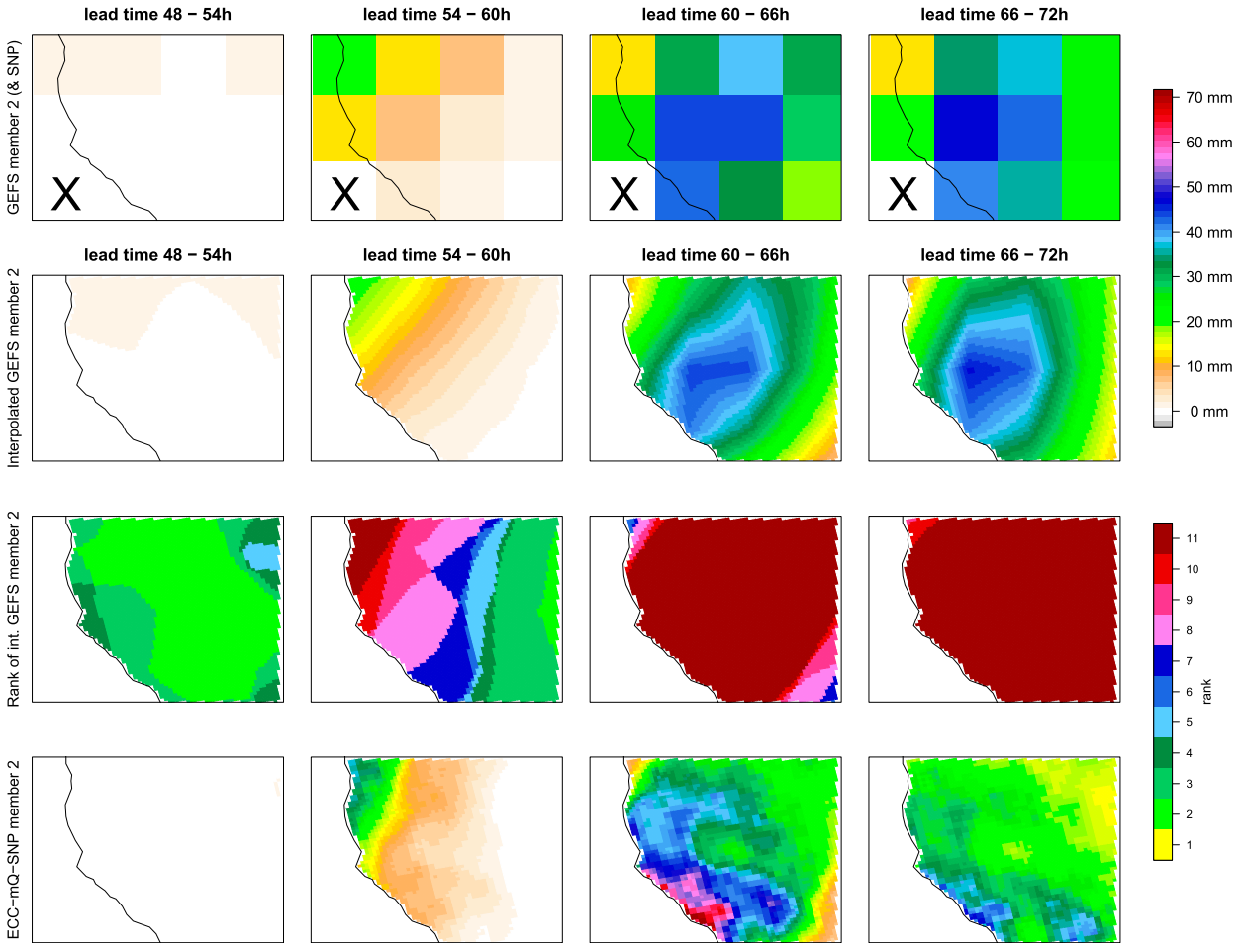

FIG. 8. (first row) Original and (second row) interpolated GEFS member 2 of the ensemble forecast initialized at 0000 UTC 17 Jan 2010, (third row) associated rank at each analysis grid point and each lead time period, and (fourth row) ECC-mQ-SNP forecast fields obtained by selecting the modified value of the predictive sample corresponding to this rank.

Figure 7 illustrates the regularization effect of the mapping $\hat{\psi}$ constructed as described above. It is clear that a linear mapping would not be able to approximate the ECC-Q point pairs sufficiently well and would therefore result in inferior sampling of the predictive distribution. The regularized linear regression spline $\hat{\psi}$, on the contrary, follows these point pairs closely except for the pairs corresponding to z_2 and z_6, which are so close together that an accurate fit would entail a steep increase of the slope of $\hat{\psi}$. The regularization results in values $\hat{z}_2$ and $\hat{z}_6$ that are also close together, which degrades the sampling of the predictive distribution (red diamonds) but eliminates the discontinuities near grid points where the ranking of two interpolated raw ensemble members changes. In other words, by controlling the gradients of the *pointwise* mapping function $\hat{\psi}$ at each analysis grid point we achieve *spatial* continuity of the mapping.

The choice of the regularization parameter $\hat{\lambda}$ is again subjective, since it trades off closeness to the CRPS-optimal ECC-Q sampling of the predictive marginal distributions (which can be quantified objectively) with a reduction in unrealistic, steep gradients in the resulting calibrated ensemble forecast fields (which at present we can only assess subjectively). In this study, we used $\hat{\lambda} = 0.5$, which resulted in sufficiently smooth ensemble forecast fields (see Fig. 8) while retaining the good sampling properties of ECC-Q (see section 4). Unlike any form of spatial smoothing, the approach described above notably modifies the ECC-Q forecast fields only in the vicinity of intersections of the interpolated raw ensemble fields which avoids the blurring of spatial detail. Since the two major changes in relation to ECC-Q are 1) the simulation of negative precipitation at forecast grid points where the raw ensemble forecasts are zero and 2) the modification of the predictive quantiles used by ECC-Q, we refer to this extension as ECC-mQ-SNP. For the practical implementation of the quantile modification via regularized linear spline

TABLE 1. Summary of the key features of the different methods for generating ensembles of space–time precipitation fields discussed in section 3.

	Optimal sampling	Represents small-scale variability	Avoids random reordering	Avoids discontinuities	Suitable for large domains	Flow-dependent space–time variability
StSS	✓	✓			✓	
MDSS-RO	✓	✓				✓
MDSS-SDA		✓	✓	✓		✓
ECC-Q	✓				✓	✓
ECC-mQ-SNP			✓	✓	✓	✓

regression, we note that the minimization in (5) is unconstrained, and the coefficients can therefore be found as the solution of a linear equation system. This technical detail is explained in appendix B.

4. Verification

Table 1 provides a summary of the key features of the different methods discussed above. Figures 2, 3, 4, 6, and 8 suggest that the ensembles of high-resolution precipitation fields generated by the MDSS-SDA or the ECC-mQ-SNP approach are physically more realistic than those generated by the StSS, MDSS-RO, or ECC-Q. Here, we study how far this translates into better performance of the resulting probabilistic forecasts.

A disadvantage of the MDSS-SDA and ECC-mQ-SNP ensembles pointed out above is that their values at each analysis grid point are not an optimal representation of the respective predictive distribution. The ECC-mQ-SNP ensemble deviates from the optimal sample due to the regression spline regularization while the MDSS-SDA algorithm generates an optimal sample at the forecast grid scale but introduces additional variability to the finescale samples during the disaggregation step. How strong is the degradation in marginal skill compared to StSS, MDSS-RO, and ECC-Q, which are all based on the same, CRPS-optimal samples? To answer this we calculate, separately for each month, CRPS skill scores relative to a climatological ensemble composed of all analyzed values (i.e., over all verification years and days within that month) at the respective analysis grid point and lead time. The results in Table 2 show that the marginal skill drops by about 0.01 for ECC-mQ-SNP and slightly more for MDSS-SDA. This loss in marginal forecast skill is not negligible, but it is still rather moderate and can be expected to be even smaller for larger ensembles. It is the price that one needs to be willing to pay for the improved spatial properties of the ensembles obtained with these methods.

In the light of the different impacts of zeros in the historic analysis ensembles or interpolated raw NWP ensembles on which the different multivariate postprocessing methods are based, verification metrics that allow one to study weather regimes with light and heavy precipitation separately are more adequate. Sample stratification comes with a number of serious caveats (Siegert et al. 2012; Lerch et al. 2017; Bellier et al. 2017b), and we therefore prefer the alternative verification strategy of studying exceedances of low, intermediate, and high thresholds; that is, we turn the analysis and ensemble forecast fields into binary fields that have the value one if the precipitation amount at this time/location exceeds the threshold, and zero otherwise. Working with binary exceedance fields is common practice in spatial forecast verification (Gilleland et al. 2009), but most of the established spatial verification methods are targeted at deterministic forecasts and are not suitable for assessing the quality of the uncertainty representation by (calibrated) ensemble precipitation forecasts. Here, we study a particular quantity that is relevant in practice and highlights the limitations of some of the multivariate postprocessing methods discussed in section 3: the fraction of all analysis grid points within the domain at which the threshold is exceeded. Roberts and Lean (2008) use a similar idea for short-range deterministic forecasts and examine at which spatial scale the predicted fraction of exceedance becomes skillful. Dey et al. (2016) extend this concept to ensembles and study spatial scales over which ensemble forecasts agree. In the present setup we are more interested in whether the different calibrated ensembles adequately represent the forecast uncertainty,

TABLE 2. CRPSSs for the (marginal) predictive distributions at each analysis grid point, aggregated over the domain, the four lead time periods, and all verification years.

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
StSS/MDSS-RO/ECC-Q	0.342	0.249	0.295	0.259	0.299	0.276	0.268	0.300
MDSS-SDA	0.336	0.238	0.283	0.247	0.286	0.263	0.257	0.287
ECC-mQ-SNP	0.339	0.238	0.287	0.248	0.292	0.267	0.260	0.291

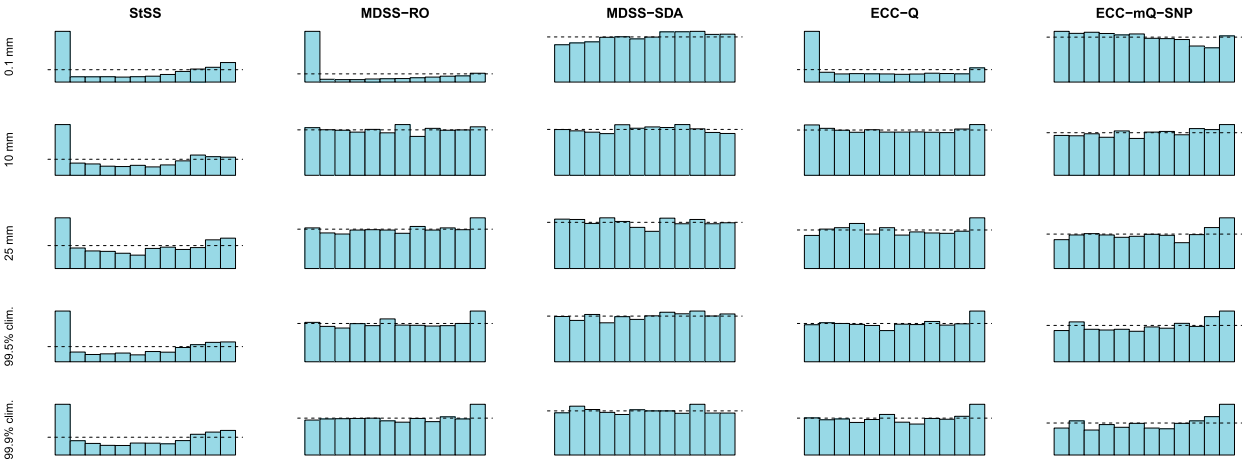

FIG. 9. Histograms of the rank of the analysis FTE within the joint sample of analysis and postprocessed ensemble FTEs for five different threshold values: 0.1 mm, 10 mm, 25 mm, the 99.5%, and the 99.9% climatological percentile at each grid point and each lead time.

and instead of a range of spatial scales we only consider the entire forecast domain (i.e., all 3262 analysis grid points) at each lead time and compare the fractions of threshold exceedance (FTEs) of analyzed and ensemble forecast fields. These FTEs can be used as a *preranks* in the sense of Gneiting et al. (2008) and Thorarinsdottir et al. (2016), that is, as projections of a multivariate quantity onto a (positive) univariate quantity that can then be evaluated by an ordinary rank histogram. If the postprocessed ensemble forecast fields represent the forecast uncertainty about the fraction of grid points within the domain that exceeds a prespecified amount of precipitation adequately, then the rank of the analysis FTE within the set consisting of all (analysis and 11 ensemble member) FTEs should assume any of the values $1, \ldots, 12$ with equal probability. If over the set of all verification cases (12 years $\times$ 8 months $\times \{28, \ldots, 31\}$ days $\times$ 4 lead times) some ranks are assumed more often than others, this can be seen as an indication for differences in the properties of the postprocessed ensemble fields and the verifying analysis fields. In the present setup with rather short accumulation periods and a domain in which there is no precipitation at all during ~40% of all days in January and ~70% of all days in May, it happens quite frequently that the FTE of the analysis and/or several ensemble members is zero. If all 12 FTEs are zero, the respective case can be omitted from the verification since it does not convey any information. If, however, at least one of the ensemble FTEs is positive, this case does inform the FTE ranking and must be used. Ties between a zero analysis FTE and zero ensemble member FTEs are then resolved at random.

Figure 9 shows the resulting FTE histograms for different threshold values. It is apparent that the StSS precipitation forecast fields differ significantly from the verifying analysis fields. The main problem is the random reordering of univariate forecast samples already discussed in connection with Fig. 2. This randomization results in very low spread of the FTEs, especially among the drier members, and this makes it rather likely that the analysis FTE is smaller than all of the ensemble FTEs. Wetter members are generally less affected by random reordering, but there is still a noticeable decrease in the spread between wet member FTEs, which results in an increased chance of the analysis FTE being the largest. The MDSS-RO and ECC-Q ensembles are also based on sample reordering, but the precipitation fields that inform the reordering (systematically selected analysis fields for MDSS-RO and interpolated raw ensemble forecasts for ECC-Q) are flow dependent, and hence in wet forecast situations there are always a sufficient number of unrandomized ranks for reordering. For the higher thresholds, the FTE histograms for MDSS-RO and ECC-Q are therefore relatively flat. In dry forecast situations, however, the flow dependence of these methods results in *more* randomization as StSS, because there is now an increased chance that there are more ensemble members/selected historic fields with zero precipitation than sample values from the marginal predictive distributions. As a result, some of the MDSS-RO/ECC-Q ensemble members have nonzero precipitation amounts at scattered analysis grid points (e.g., MDSS-RO members 3, 5, and 6 in online supplement B) and result in poor calibration at the 0.1-mm threshold. The MDSS-SDA method avoids random reordering at the analysis grid scale entirely. For the coarse-scale fields (second row in Fig. 4) it can still occur that too many values of the upscaled historic analysis fields are zero. In this case the values of the corresponding MDSS-SDA ensemble member are also zero at all

analysis grid points associated with this forecast grid cell, even if the coarse-scale predictive sample value at this grid cell suggests a small nonzero precipitation amount. This explains the subtle dry bias of the MDSS-SDA ensembles at the 0.1-mm threshold, but overall this approach yields the best representation of the FTE over the domain. The ECC-mQ-SNP also avoids random reordering by simulating negative precipitation at the forecast grid scale and interpolating it (along with the nonzero precipitation amounts from the raw ensemble forecasts) to the analysis grid points. At the higher thresholds, the FTE histograms look the same as for ECC-Q, but the miscalibration of the ECC-Q ensembles at the 0.1-mm threshold is largely rectified. There is still some departure from a flat histogram, which is likely a result of our ad hoc simulation procedure for the negative precipitation amounts. This procedure seems to generally capture the strength of spatial covariability, but it is unable to perfectly mimic the nature of real precipitation fields. We leave the development of more sophisticated and statistically principled simulation methods for future research.

The FTE histograms in Fig. 9 are useful to diagnose shortcomings of the different methods to generate ensembles of precipitation forecast fields, but they do not permit a quantitative assessment of forecast performance. This can be done though by comparing continuous ranked probability skill scores (CRPSSs) of the different calibrated ensemble FTEs relative to the FTEs of a climatological ensemble. In addition to reliability, this also assesses sharpness of the ensemble precipitation fields, that is, how well the ensemble FTEs are concentrated around the analysis FTE. The results in Table 3 confirm that StSS fares significantly worse than the other methods across all thresholds. Interestingly, we now see an advantage of the two MDSS methods over the ECC methods at the high thresholds, especially for the threshold values relative to climatology, where terrain-related spatial differences are largely removed. This might be caused by the inability of ECC to account for subgrid-scale variability, which causes the wettest members to exceed a high precipitation threshold somewhat less often than the wettest MDSS-SDA members.

Another aspect of precipitation fields that is particularly important for hydrological applications is that spatially and temporally accumulated precipitation amounts are represented correctly. Comparisons of different multivariate postprocessing techniques in that regard have been performed, for example, by Scheuerer et al. (2017) and Wu et al. (2018). Accumulations over space and time, however, are more of a coarse-scale quantity, and in the present setup it is therefore more interesting to study genuine subgrid-scale aspects like the maximum precipitation amount over a certain subdomain and a certain time period. In practice, such a subdomain could be some administrative unit for which severe weather warnings such as a flash flood warning are issued. Here, we use each GEFS grid cell as a subdomain, and we study the maximum over all associated finescale grid points and all four lead time periods. Asking whether the observed spatiotemporal maximum exceeds 0.1 mm of precipitation is equivalent to asking whether there is measurable precipitation anywhere within the subdomain during any of the four 6-h periods. Likewise, when studying exceedance of 25 mm, we are interested if during any of the 6-h periods there is at least one analysis grid point within the subdomain where an elevated amount of precipitation is reported. Studying these subgrid maxima is also interesting because the raw ensemble forecasts do not provide any subgrid-scale information, and it is therefore not clear if ECC-based downscaling can provide good forecasts of this quantity.

TABLE 3. CRPSSs for the fractions of threshold exceedance for different thresholds, aggregated over the four lead time periods, all months, and all verification years.

	0.1 mm	10 mm	25 mm	99.5% climatology	99.9% climatology
StSS	0.338	0.149	0.065	0.120	0.088
MDSS-RO	0.439	0.225	0.126	0.206	0.161
MDSS-SDA	0.478	0.225	0.126	0.207	0.160
ECC-Q	0.462	0.218	0.116	0.196	0.145
ECC-mQ-SNP	0.473	0.211	0.122	0.191	0.142

Figure 10 shows Brier skill scores of subgrid-scale maximum precipitation forecasts obtained with the four different approaches where skill is again relative to climatological forecasts. This time, however, climatological exceedances are calculated with all analyzed fields during a 91-day window centered over the middle of the respective month in order to reduce the variability of the estimated climatological frequencies of exceeding the 25 mm threshold. The scores reinforce the conclusions obtained with the FTE-based metrics considered above. The MDSS-SDA algorithm yields a significantly better representation of subgrid-scale maximum precipitation than the StSS method and attains much better skill scores at all thresholds. The poor performance of StSS can again be explained by the randomization of predictive samples when the associated historic cases used for reordering are zero. The detrimental effects of this common practice appear quite markedly here since the distribution of the spatiotemporal maximum is rather sensitive to the assumption of weak spatial dependence entailed by the random reordering for a subset of ensemble members and subsets of the spatial domain.

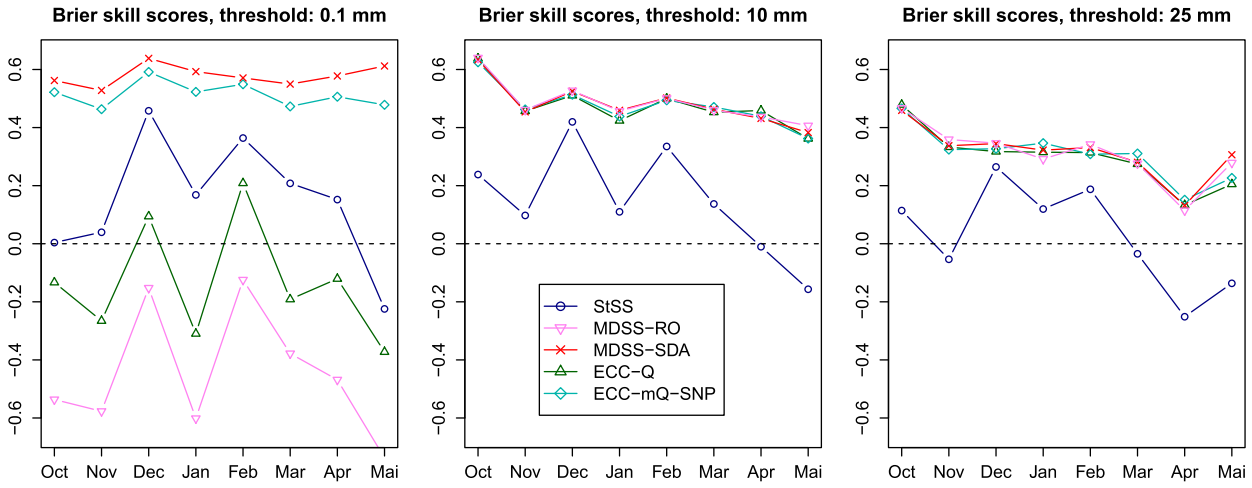

FIG. 10. Brier skill scores for maximum, analysis-scale 6-h precipitation amounts over each GEFS grid cell and four lead time periods.

The same cause explains the poor performance of the MDSS-RO and the ECC-Q method at the 0.1-mm threshold as discussed above in connection with the FTE histograms. Interestingly, the two ECC methods perform as well as MDSS-RO and MDSS-SDA at the higher thresholds, which suggests that the ECC ensembles are able to represent the predictive distribution of subgrid-scale maxima quite well despite the lack of information about subgrid-scale variability in the raw ensemble. The discontinuities in the MDSS-RO and ECC-Q ensembles noted in Figs. 3 and 6 do not seem to have a significant negative effect on the distribution of subgrid-scale maximum precipitation amounts, but neither does the suboptimal sampling of MDSS-SDA and ECC-mQ-SNP; results for all four methods are very similar at the higher thresholds, and different verification metrics might have to be devised to appreciate the more realistic appearance of the MDSS-SDA and ECC-mQ-SNP ensemble precipitation fields.

5. Discussion

We have reviewed two established methods, the Schaake shuffle (StSS) and ECC-Q, and the recently proposed minimum divergence Schaake shuffle (MDSS) for modeling space–time variability in postprocessed ensemble precipitation forecasts. A computationally efficient variant (MDSS-SDA) of the purely reordering-based MDSS-RO implementation has been proposed that allows one to generate high-resolution precipitation fields based on coarser-resolution numerical weather predictions while ensuring that precipitation fields generated with this new approach are physically realistic. We also discussed an extension (ECC-mQ-SNP) of the ECC-Q approach that retains its good sampling properties while eliminating discontinuities in the ECC-Q ensemble forecast fields and avoiding random reordering of predictive samples in the situation where most or all of the raw ensemble forecasts are zero. A new diagnostic tool, the fraction of threshold exceedance (FTE) histogram, has been proposed and was used to demonstrate that the MDSS-SDA and ECC-mQ-SNP ensembles represent the fractional areal coverage of precipitation exceeding a predefined threshold better than the StSS ensembles for all thresholds and better than the MDSS-RO and ECC-Q ensembles at the lowest threshold. These conclusions were confirmed by quantitative verification of probabilistic, subgrid-scale maximum precipitation forecasts derived from the different postprocessed ensemble precipitation fields. All verification results presented in section 4 were obtained with 11-member postprocessed ensembles. For the two ECC methods this ensemble size is stipulated by the size of the GEFS ensemble, but the StSS and MDSS ensembles are not subject to that restriction and were constructed to be of the same size only to permit a direct comparison. Results by Wilks (2015), Scheuerer et al. (2017), and Wu et al. (2018) suggest that larger ensembles could improve the scores for StSS, MDSS-RO, and MDSS-SDA reported in section 4, but we note that the shortcomings (random reordering, discontinuities) of the StSS and MDSS-RO method pointed out in section 3 still exist with larger ensembles.

The poor performance of StSS in all and MDSS-RO and ECC-Q in dry weather situations was attributed to the challenges that occur when the raw ensemble forecasts (for ECC) or the historic fields (for StSS and MDSS-RO) contain more zeros than the calibrated forecast samples they are supposed to reorder. MDSS-SDA avoids this situation and attained the overall best verification scores in our study. Its biggest limitation compared

to the StSS and the ECC methods is that it is not useful in applications where the spatial domain of interest is large. Even with a long archive of historic analyses, it becomes increasingly difficult to find a set of historic analysis fields where the marginal distributions at all lead times and all grid points in a large domain resemble those of the calibrated univariate forecast distributions. The implementation presented here alleviates this problem somewhat in that it reduces the spatial dimension from the number of analysis scale grid points to the number of forecast grid points within the domain. It is therefore valuable in combination with a high-resolution, distributed hydrological forecast system over a river basin of limited size like the Russian River watershed considered in our case study. For large basins, the ECC-mQ-SNP method proposed in section 3f is probably the most promising approach to modeling space–time variability of precipitation forecast fields. While it cannot represent subgrid variability as good as an analysis-based modeling approach, this technique has no limitations with regard to the size of the spatial domain to which it is applied, and that makes it a rather universal, multivariate modeling approach. Our proposed strategy to avoid random reordering—simulation of negative precipitation—is rather ad hoc, and further research is required to develop a statistically more principled way to simulate negative precipitation and further improve the implied model for the spatial structure of light precipitation amounts.

The verification performed in this paper has focused on spatial and subgrid-scale properties of the postprocessed ensemble forecast fields. In a follow-up project we will evaluate their performance when used as meteorological forcings of the U.S. National Water Model.

Acknowledgments. We are grateful for the numerous constructive suggestions by two anonymous reviewers, which led to significant improvements in the presentation of this work. This research was supported by a grant from the NOAA/NWS Research to Operations (R2O) Joint Technology Transfer Initiative (JTTI), Award NA17OAR4320101 and by a funding agreement between NOAA/ESRL/PSD and NOAA/NWS/MDL on the development and transfer of advanced statistically postprocessed probabilistic ensemble guidance.

APPENDIX A

Construction of the Basis Functions in Fig. 5

The basis functions $\varphi_1, \ldots, \varphi_M$ are intended to induce a smooth function $\eta^{(k)}$ when linearly combined as in (2) while having a support radius ρ that is as small as possible, so that they primarily affect the multiplicative adjustment within the associated grid cell. There are numerous ways to define basis functions that fulfill these two criteria, and here we describe a construction with which we obtained good results. Using geographic coordinates (ϕ, λ), we first define functions $\vartheta_1, \ldots, \vartheta_M$ via

$$\vartheta_m(\cdot) := \left[1 - \left(\frac{\phi_m - \phi_\cdot}{\rho}\right)^3\right]_+^3 \left[1 - \left(\frac{\lambda_m - \lambda_\cdot}{\rho}\right)^3\right]_+^3, \tag{A1}$$

where $(\cdot)_+ = \max(\cdot, 0)$. This way, each $\vartheta_m(\cdot)$ is the Cartesian product of one-dimensional tricube kernels (see lower-left corner in Fig. 5) in the meridional and zonal direction. This choice was made because tricube kernels are smooth but have a relatively flat top, which avoids strong spikes of $\eta^{(k)}$. Moreover, they are compactly supported, that is, they are zero outside their support radius ρ. The basis functions $\varphi_1, \ldots, \varphi_M$ are then obtained by normalizing these functions at each analysis grid point

$$\varphi_m(\cdot) := \frac{\vartheta_m(\cdot)}{\sum_{m'=1}^{M} \vartheta_{m'}(\cdot)}, \tag{A2}$$

thus achieving the desirable property that a constant adjustment function $\eta^{(k)}$ is obtained when all $\alpha_{km'}$ are identical. The function φ_m is equal to zero outside the square with edge length 2ρ, centered over forecast grid point m. In this study we chose ρ equal to 1.5 times the forecast grid resolution, which we found to be a good trade-off between localization and smoothness of the resulting function $\eta^{(k)}$.

APPENDIX B

Efficient Minimization of Eq. (5)

Define

$$\mathbf{A} = \begin{bmatrix} 1 & z_{\pi(\tilde{n}_0)} & \left(z_{\pi(\tilde{n}_0)} - z_{\pi(\tilde{n}_0+1)}\right)_+ & \cdots & \left(z_{\pi(\tilde{n}_0)} - z_{\pi(K-1)}\right)_+ \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_{\pi(K)} & \left(z_{\pi(K)} - z_{\pi(\tilde{n}_0+1)}\right)_+ & \cdots & \left(z_{\pi(K)} - z_{\pi(K-1)}\right)_+ \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} \tilde{z}_{\pi(\tilde{n}_0)} \\ \cdots \\ \tilde{z}_{\pi(K)} \end{pmatrix},$$

let $\mathbf{D}$ be the diagonal matrix with diagonal entries 0, 0, $\max(\tilde{z}_{\pi(\tilde{n}_0+1)}, 1)$, …, $\max(\tilde{z}_{\pi(K-1)}, 1)$, and denote by $\boldsymbol{\gamma}$ the vector of spline coefficients. The target function in (5) can then be written as

$$T\left(\gamma_0, \ldots, \gamma_{K-\tilde{n}_0}\right) = (\mathbf{A}\boldsymbol{\gamma} - \mathbf{b})^{\mathrm{T}}(\mathbf{A}\boldsymbol{\gamma} - \mathbf{b}) + \hat{\lambda}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{D}\boldsymbol{\gamma}.$$

This is a quadratic form in $\boldsymbol{\gamma}$ and some basic calculus shows that the minimizers of this quadratic form can be found as the solution to the linear equation system $(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \hat{\lambda}\mathbf{D})\boldsymbol{\gamma} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$.

REFERENCES

Bellier, J., G. Bontron, and I. Zin, 2017a: Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resour. Res.*, **53**, 10 085–10 107, https://doi.org/10.1002/2017WR021245.

——, I. Zin, and G. Bontron, 2017b: Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Mon. Wea. Rev.*, **145**, 3529–3544, https://doi.org/10.1175/MWR-D-16-0487.1.

——, ——, and ——, 2018: Generating coherent ensemble forecasts after hydrological postprocessing: Adaptations of ECC-based methods. *Water Resour. Res.*, **54**, 5741–5762, https://doi.org/10.1029/2018WR022601.

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524, https://doi.org/10.1175/WAF-D-12-00062.1.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2008: Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.*, **2**, 1170–1193, https://doi.org/10.1214/08-AOAS203.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Bröcker, J., 2012: Evaluating raw ensembles with the continuous ranked probability score. *Quart. J. Roy. Meteor. Soc.*, **138**, 1611–1617, https://doi.org/10.1002/qj.1891.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu, 1995: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208, https://doi.org/10.1137/0916069.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901, https://doi.org/10.1175/2009MWR3187.1.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Dey, S. R. A., N. M. Roberts, R. S. Plant, and S. Migliorini, 2016: A new method for the characterization and verification of local spatial predictability for convective-scale ensembles. *Quart. J. Roy. Meteor. Soc.*, **142**, 1982–1996, https://doi.org/10.1002/qj.2792.

Flowerdew, J., 2014: Calibrating ensemble reliability whilst preserving spatial structure. *Tellus*, **66A**, 22662, https://doi.org/10.3402/tellusa.v66.22662.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, 2008: Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *TEST*, **17**, 211–264, https://doi.org/10.1007/s11749-008-0114-x.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, https://doi.org/10.1175/MWR-D-18-0147.1, in press.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

——, 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 data set. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

Herman, G. R., and R. S. Schumacher, 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.

Herr, H. D., and R. Krzysztofowicz, 2005: Generic probability distribution of rainfall in space: The bivariate model. *J. Hydrol.*, **306**, 234–263, https://doi.org/10.1016/j.jhydrol.2004.09.011.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

Kleiber, W., A. E. Raftery, and T. Gneiting, 2011: Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Amer. Stat. Assoc.*, **106**, 1291–1303, https://doi.org/10.1198/jasa.2011.ap10433.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Stat. Sci.*, **32**, 106–127, https://doi.org/10.1214/16-STS588.

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.*, **134**, 2029–2050, https://doi.org/10.1002/qj.334.

R Core Team, 2017: R: A language and environment for statistical computing. R Foundation for Statistical Computing, https://www.R-project.org/.

Roberts, N., and H. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Robertson, D. E., D. L. Shresta, and Q. J. Wang, 2013: Postprocessing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.*, **17**, 3587–3603, https://doi.org/10.5194/hess-17-3587-2013.

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.*, **140**, 874–888, https://doi.org/10.1175/MWR-D-11-00062.1.

Schaake, J., and Coauthors, 2007: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci. Discuss.*, **4**, 655–717, https://doi.org/10.5194/hessd-4-655-2007.

Schefzik, R., 2016: A similarity-based implementation of the Schaake shuffle. *Mon. Wea. Rev.*, **144**, 1909–1921, https://doi.org/10.1175/MWR-D-15-0227.1.

——, T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, https://doi.org/10.1214/13-STS443.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

——, and T. M. Hamill, 2015: Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

——, ——, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatio-temporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, https://doi.org/10.1002/2016WR020133.

Siegert, S., J. Bröcker, and H. Kantz, 2012: Rank histograms of stratified Monte Carlo ensembles. *Mon. Wea. Rev.*, **140**, 1558–1571, https://doi.org/10.1175/MWR-D-11-00302.1.

Sloughter, J. M., 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.

Stauffer, R., N. Umlauf, J. W. Messner, G. Mayr, and A. Zeileis, 2017: Ensemble post-processing of daily precipitation sums over complex terrain using censored, high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, https://doi.org/10.1175/MWR-D-16-0260.1.

Thorarinsdottir, T. L., M. Scheuerer, and C. Heinz, 2016: Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.*, **25**, 105–122, https://doi.org/10.1080/10618600.2014.977447.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Whan, K., and M. Schmeits, 2018: Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Mon. Wea. Rev.*, https://doi.org/10.1175/MWR-D-17-0290.1, in press.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, https://doi.org/10.1002/met.134.

——, 2015: Multivariate ensemble Model Output Statistics using empirical copulas. *Quart. J. Roy. Meteor. Soc.*, **141**, 945–952, https://doi.org/10.1002/qj.2414.

Wu, L., D.-J. Seo, J. Demargne, J. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *J. Hydrol.*, **399**, 281–298, https://doi.org/10.1016/j.jhydrol.2011.01.013.

——, Y. Zhang, T. Adams, H. Lee, Y. Liu, and J. Schaake, 2018: Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *J. Hydrometeor.*, **19**, 575–598, https://doi.org/10.1175/JHM-D-17-0054.1.

Zhang, Y., L. Wu, M. Scheuerer, J. Schaake, and C. Kongoli, 2017: Comparison of probabilistic quantitative precipitation forecasts from two postprocessing mechanisms. *J. Hydrometeor.*, **18**, 2873–2891, https://doi.org/10.1175/JHM-D-16-0293.1.