Comparative Evaluation of Three Schaake Shuffle Schemes in Postprocessing GEFS Precipitation Ensemble Forecasts

LIMIN WU

Lynker Technologies, and Office of Water Prediction, Silver Spring, Maryland

YU ZHANG

The University of Texas at Arlington, Arlington, Texas

THOMAS ADAMS

TerraPredictions, Blacksburg, Virginia

HAKSU LEE AND YUQIONG LIU

LEN Technologies, and Office of Water Prediction, Silver Spring, Maryland

JOHN SCHAAKE

Consultant, Annapolis, Maryland

(Manuscript received 31 March 2017, in final form 7 October 2017)

ABSTRACT

Natural weather systems possess certain spatiotemporal variability and correlations. Preserving these spatiotemporal properties is a significant challenge in postprocessing ensemble weather forecasts. To address this challenge, several rank-based methods, the Schaake Shuffle and its variants, have been developed in recent years. This paper presents an extensive assessment of the Schaake Shuffle and its two variants. These schemes differ in how the reference multivariate rank structure is established. The first scheme (SS-CLM), an implementation of the original Schaake Shuffle method, relies on climatological observations to construct rank structures. The second scheme (SS-ANA) utilizes precipitation event analogs obtained from a historical archive of observations. The third scheme (SS-ENS) employs ensemble members from the Global Ensemble Forecast System (GEFS). Each of the three schemes is applied to postprocess precipitation ensemble forecasts from the GEFS for its first three forecast days over the mid-Atlantic region of the United States. In general, the effectiveness of these schemes depends on several factors, including the season (or precipitation pattern) and the level of gridcell aggregation. It is found that 1) the SS-CLM and SS-ANA behave similarly in spatial and temporal correlations; 2) by a measure for capturing spatial variability, the SS-ENS outperforms the SS-ANA, which in turn outperforms the SS-CLM; and 3), overall, the SS-ANA performs better than the SS-CLM. The study also reveals that it is important to choose a proper size for the postprocessed ensembles in order to capture extreme precipitation events.

1. Introduction

Because of their ability to depict forecast uncertainties and improve forecast accuracy, ensemble weather and climate forecasts have seen wide applications in hydrologic forecasting, water resources management, and emergency preparedness (Georgakakos et al. 1998; Ajami et al. 2008). Despite substantial improvements made in model physics and data assimilation in recent decades, raw ensemble forecasts from dynamical models remain subject to large systematic and random errors (Hamill and Whitaker 2006). Moreover, downscaling is often required for these forecasts in downstream applications such as hydrologic forecasting. To address these shortcomings, various techniques have been developed and evaluated over the years to postprocess ensemble forecasts (Raftery et al. 2005;

Corresponding author: Limin Wu, limin.wu@noaa.gov

DOI: 10.1175/JHM-D-17-0054.1

Hamill and Whitaker 2006; Schaake et al. 2007; Wu et al. 2011; Cui et al. 2012; Scheuerer and Hamill 2015; Yang et al. 2017).

Natural weather systems possess certain spatiotemporal variability and correlations. Preserving these spatiotemporal properties is necessary for applications in ensemble streamflow predictions and particularly critical for flood forecasting, since the structure of a storm and the direction in which the storm passes over a watershed could determine the timing and magnitude of flood peaks. Characterizing spatiotemporal properties involves multivariate statistical modeling. One approach to ensemble postprocessing in a multivariate setting is by extending conventional parametric methods for univariate ensemble postprocessing. These extensions, however, require the estimation of large numbers of parameters (Gel et al. 2004; Berrocal et al. 2007; Berrocal and Raftery 2008; Pinson 2012; Schuhen et al. 2012), which can severely limit their applicability due to computational constraints. Another approach, instead of using a full parametric modeling strategy, employs an empirical copula framework for postprocessing. This approach has gained popularity in recent years and attracted significant research efforts (Clark et al. 2004; Schaake et al. 2007; Voisin et al. 2010; Robertson et al. 2013; Schefzik et al. 2013; Wilks 2015; Schefzik 2016; Scheuerer et al. 2017). It effectively addresses the dimensionality problem arising from extending conventional parametric methods. In this approach, a rank-based shuffling technique plays a key role. The technique relies on the construction of rank structures from a certain source of forecasts or observations. The precise meaning of a rank structure will be given in section 2b. The original concept of the technique, known as the Schaake Shuffle, is published in Clark et al. (2004).

With the original Schaake Shuffle method, rank structures are provided by observations collected from the days surrounding the forecast date, not only in the same year, but also across the years in a historical archive. The rationale behind the method is that through the use of this rank structure, the postprocessed ensemble members may acquire the spatiotemporal coherence of natural weather regimes. The method, however, receives criticism for not accounting for the current atmospheric conditions. To address this issue, several variants of the Schaake Shuffle have been proposed and evaluated (Schefzik et al. 2013; Vrac and Friederichs 2015; Schefzik 2016; Scheuerer et al. 2017). Among these, Schefzik (2016) developed a similarity-based shuffling method (termed SimSchaake therein), following the suggestion of Clark et al. (2004). This method uses a subset of historical dates, for which the predicted atmospheric conditions resemble that of the present, for constructing rank structures. Schefzik (2016) applied this method to surface temperature and showed its effectiveness in improving the overall performance of the postprocessed ensemble forecasts. Another variant, implemented in the ensemble copula coupling (ECC; Schefzik et al. 2013) framework, makes use of ensemble forecasts for rank structure construction. This method accounts for the current atmospheric conditions through the predictive capability of the dynamical models. In addition, since the raw ensemble forecasts from dynamical models would resemble observations predicted to a certain degree, the rank structure derived would reflect the spatiotemporal coherence of natural weather regimes.

An empirical copula (EC) interpretation of the original Schaake Shuffle is given in Schefzik et al. (2013). The foundation of the copula theory is the Sklar's theorem (Sklar 1973, Theorem 1), which states that any multivariate distribution equals a copula constructed from the marginals of the distribution. The EC can be thought of as an approximation to the true copula. The EC interpretation may be rephrased as follows. In postprocessing an ensemble forecast for multiple locations and lead times, a univariate forecast probabilistic distribution can be obtained for each of the forecast points. Then, a dependence structure for these univariate distributions needs to be specified. Ideally, the dependence structure chosen in the postprocessing can yield postprocessed ensembles that are consistent with the weather and climate in terms of spatiotemporal properties. To this end, the Schaake Shuffle seeks to construct ECs from historical observations. An EC thus constructed possesses a dependence structure. This dependence structure is then applied to the EC derived from the forecast distributions by the rank-based shuffling procedure of the Schaake Shuffle. In this sense, the Schaake Shuffle can be considered an empirical copula technique.

The Schaake Shuffle method and its variants discussed above represent three general ways of obtaining rank structures. Clearly, they share a common framework: a rank structure provided by a certain source and a shuffling procedure to impose the rank structure on the processed ensemble members in a multivariate setting. In this work, we consider these three types of rank structures with specific implementations: 1) the rank structure provided by climatological observations near the forecast date, 2) one provided by analogs derived from historical observations using a particular similarity criterion, and 3) one provided by ensemble forecasts from NOAA's Global Ensemble Forecast System (GEFS). For brevity, these schemes will be referred to as SS-CLM, SS-ANA, and SS-ENS, respectively, and collectively called SS schemes hereinafter.

While assessments by Schefzik et al. (2013) and Schefzik (2016) point to performance gains for their respective approaches, these assessments are limited in spatial domain and evaluation scope and are not specific for precipitation forecasts, whose spatiotemporal structures are critical in the prediction of flooding events. In this work, we perform a detailed comparison of the SS-CLM, SS-ANA, and SS-ENS schemes in postprocessing precipitation forecasts. This effort complements the limited scope of prior studies by focusing on a large spatial domain in the eastern United States over a relatively long time period. The performance of the SS schemes is evaluated using various measures, including correlation measures and other conventional ensemble verification measures such as the Brier skill score (BSS). Evaluation of the postprocessed ensembles at various spatial scales is needed for many downstream applications, such as hydrologic streamflow prediction. To this end, the effectiveness of the SS schemes is also assessed at various levels of spatial aggregation. Additionally, spatial similarities between precipitation fields of the observed and forecast are assessed by employing the Baddeley's delta metric (Gilleland 2011). Last, the capability of the SS schemes for correctly distributing precipitation over the study domain climatologically is compared.

The rest of this paper is organized as follows. Section 2 describes the statistical procedures for postprocessing gridded GEFS precipitation ensemble forecasts. Section 3 describes the study area, datasets, and methods used in evaluating the statistical procedures. Section 4 presents the outcomes. Section 5 discusses limitations of this study. Section 6 summarizes and discusses the findings.

2. Postprocessing procedure

The postprocessing procedure employed here is an adaptation of the one used in the Meteorological Ensemble Forecast Processor (MEFP), developed by the NWS Office of Water Prediction (OWP), formerly the Office of Hydrologic Development. This adaptation is a direct response to the need of calibrated short- and medium-range ensemble weather forecasts in high-resolution grids to serve the National Water Model at the OWP. The MEFP is a basin-based system. It can produce calibrated precipitation and temperature ensemble forecasts at spatial scales of river drainage areas and temporal scales from 6 h up to 3 months, for forecast periods ranging from a few days to about 9 months (Demargne et al. 2014). The OWP has conducted a number of performance evaluations for the MEFP by both project developers and field users through event-based evaluations and statistical validation (Brown et al. 2014a,b). The evaluations include 1) the use of quantitative precipitation forecasts (QPFs) and GEFS reforecasts as input sources for postprocessing, 2) sensitivity studies to assess a number of aspects of the MEFP, and 3) a streamflow ensemble study using MEFP-processed ensembles as forcing data. These evaluations indicate that, overall, the MEFP is capable of producing reliable precipitation and temperature ensembles and preserving the forecast skill in the input forecast sources. It is worth noting that the MEFP can be deemed as an ensemble processor for irregular grids as it can be used to treat a group of forecast drainage areas with various sizes.

a. General procedure

The postprocessing procedure employed here consists of five main steps.

1) For a given location and forecast start time, select a spatial area and temporal interval in the forecast period for postprocessing. The start time, end time, and duration of the time interval is predefined. The area for the location is also predefined. Collectively, the above quantities will be referred to as a space–time forecast point hereinafter.
2) Archived historical forecast data or reforecast data need to be prepared for calibrating the statistical models to be used. The forecasts in these datasets need to be paired with the corresponding observations. To increase the sample sizes for calibration, data points may be pooled temporally and spatially over the space surrounding the given space–time forecast point. Spatially averaged and temporally accumulated values of the observed and forecast variables are obtained.
3) For a given space–time forecast point, the relationship of the forecast (single valued) and observed is modeled by a mixed-type bivariate distribution, with its continuous component modeled by the meta-Gaussian distribution (MGD; Kelly and Krzysztofowicz 1997; Herr and Krzysztofowicz 2005; Wu et al. 2011). This mixed-type bivariate distribution will be referred to as MMGD throughout this paper. At this step, parameters of the bivariate model are estimated. For a given forecast, a random sample can be drawn from the conditional distribution of the bivariate model. This is the step that provides the mechanism for error correction and uncertainty quantification for individual space–time forecast points.
4) The sample points obtained from the above step are arranged using the Schaake Shuffle method (or a variant).

5) Each ensemble member is disaggregated at a finer spatial scale and/or temporal scale if needed. This is necessary if the downstream application operates on a finer grid.

An analogy can be drawn between our postprocessing approach outlined above, referred to as MMGD-SS hereinafter, and the ECC approach (Schefzik et al. 2013). The ECC has its theoretical basis in the theory of multivariate empirical copulas. As a general ensemble postprocessing strategy, the ECC uses the rank structure provided by ensemble forecasts to specify the dependence structure of the multivariate empirical copula. The ECC consists of three steps. First, each of the marginal variables of the copula is treated individually, using a certain statistical postprocessing technique, such as ensemble Bayesian model averaging (Raftery et al. 2005). Then, a sample of a predefined size is drawn from each postprocessed predictive distribution. After that, the sampled values are arranged using an ordering procedure of the SS-ENS type. It can be recognized that the three steps of the ECC are also present in the MMGD-SS approach in a similar way. Therefore, the MMGD-SS approach also fits into the empirical copula framework. A distinct rank structure used in the MMGD-SS will result in a distinct dependence structure for the empirical copula.

b. *Schaake Shuffle*

A rank structure obtained from an SS scheme can be viewed in different ways. Perhaps it is convenient to think of it as a set of tables, each consisting of columns of ranks. Each column in such a table corresponds to an ensemble obtained at a space–time forecast point. A rank structure can be derived from various sources. The following describes in detail how a rank structure is constructed.

1) THE SS-CLM SCHEME

The rank-structure construction for this scheme can be illustrated by an example. Suppose we have a record of historical observed data of 50 years for a spatial domain. For a given spatial forecast point, the observations can be arranged as a table with each year occupying a row. The rows are ordered chronologically. The columns are also ordered chronologically from 1 January to 31 December. We thus obtain a table of 50 rows and 366 columns (assuming a 24-h time step for the data). Given a particular 24-h forecast for the spatial location, we can find a column (indexed with day of year) in the table corresponding to the lead time of the forecast. Each value of the column has a rank (the smallest value is ranked 1). Therefore, this table of observations can be mapped to a table of ranks. Collectively, all such rank tables obtained for the spatiotemporal domain constitute a rank structure for the SS-CLM scheme. Certainly, placing these tables into a three-dimensional array gives us a different view of the rank structure. For a given space–time forecast point, the scheme arranges the members of an ensemble according to the ranks of the rank structure for that space–time forecast point. We note here that in order for the SS-CLM to work properly, the same historical years must be used for each location in the forecast spatial domain.

2) THE SS-ANA SCHEME

Analogs of precipitation events may be obtained in many ways. There exists a large body of literature concerning measurements of similarity between weather objects in the area of spatial weather forecast verification (Zepeda-Arce et al. 2000; Venugopal et al. 2005; Ebert 2008, 2009; Clark et al. 2010; Marzban and Sandgathe 2010; Gilleland 2011), wherein a variety of methods can be used to search for analogs. Among them, an early method termed upscaling, perhaps best known in the category of neighborhood methods (Ebert 2009), is chosen and used in this SS scheme. The method first averages current and past forecasts to coarser space–time resolutions and then applies a certain similarity criterion to find analogs for the current forecast. Here, the method is implemented with the RMSE between the current forecast and the past forecasts on aggregated areas. We note here that the similarity criterion employed here is different than the one used in Schefzik (2016), where empirical standard deviation is involved. Specifically, rank-structure construction for this scheme proceeds as follows. First, for a given gridded ensemble forecast over a spatial domain at a temporal forecast point (or lead time), ensemble members are averaged over the entire domain to obtain an ensemble-mean field. Then, the ensemble-mean field is compared with a sequence of ensemble-mean fields similarly obtained from historical forecasts or retrospective forecasts. The comparison is done using the RMSE on upscaled values of aggregated areas with prescribed sizes. A prespecified number of ensemble-mean fields associated with the smallest RMSE values can then be selected from the past forecasts. Each of these fields has a corresponding observed field and an associated date. Thus, a sequence of historical dates with their associated observations is obtained. Recall that in the SS-CLM scheme, a rank structure can be constructed from a sequence of historical dates and observations. Here, as the last step of the SS-ANA, we use that procedure of the SS-CLM to obtain rank structures.

3) THE SS-ENS SCHEME

A rank structure for this scheme can be obtained as follows. Suppose a gridded ensemble forecast is issued. For a given space–time forecast point, we have a fixed number of ensemble members: a control member, perturbed member 1, ... , perturbed member n. These numbers can be placed into a list in the stated order. With the numbers replaced by their respective ranks, we obtain a sequence of ranks. By putting all such sequences into a table column-wise, we obtain a rank structure of the SS-ENS type.

The shuffling part of the SS schemes is straightforward. The effect of the shuffling can be illustrated by the following hypothetical example. Suppose a rank structure is provided by a source, and the final ensemble has five members. For a space–time forecast point, the values of the source, the ranks of these values, and the values of the final output ensemble are given in Table 1. The first column of this table shows the values of the source in a predetermined order. The second column shows the corresponding ranks for these values, taken from the rank structure for the space–time forecast point. Here, the smallest value is ranked 1. The order of the ensemble members coming out from the MMGD model can be arbitrary, usually in ascending or descending order. After the shuffling, the order of the ensemble members (column 3) follows that of the source values; that is, the ensemble members and the source values have common ranks (column 2).

A known problem with the SS schemes when applied to precipitation is that in the rank-structure construction, there can exist a large number of zero precipitation values for space–time points of a small scale, resulting in many tied ranks. In the case of a heavy precipitation event being predicted, some ensemble members with positive values will be ranked randomly as a result of tied matching ranks. This is evidently true for the SS-CLM scheme, as the number of tied ranks is related to the probability of precipitation (PoP) at the space–time point. For a PoP at 0.3, for example, there will be, on average, 35 tied ranks for an ensemble of 50 members. Intuitively, this problem may be mitigated by using the SS-ENS or SS-ANA scheme. As an example, consider the following scenario. At a space–time forecast point, an ensemble forecast from a dynamical model predicts a high probability of heavy precipitation, that is, members of large values constitute a large fraction of the ensemble. In this scenario, more unique ranks would be provided by the SS-ENS scheme than the SS-CLM, since the number of unique ranks provided by the SS-CLM is constrained by the PoP for the space–time forecast point. As a result of the heavy precipitation prediction from the dynamical model, the ensemble coming from the MMGD model should also have a large fraction of positive members. Therefore, the SS-ENS scheme would result in fewer tied ranks than the SS-CLM scheme.

TABLE 1. An illustration of the SS schemes.

Source (mm)	Rank	Ensemble (mm)
0.7	4	7.0
0.2	2	0.3
0.4	3	3.7
0.0	1	0.1
1.9	5	9.5

Random ranking in the SS schemes may lead to unrealistic spatial and temporal correlations for the ensembles. In the appendix, we give an analysis on the Spearman's rank correlation in the tied-rank situation, illustrated with the SS-CLM scheme. The analysis also applies to the other two schemes. The analysis shows that the presence of tied ranks in the rank structures can compromise the effectiveness of the SS schemes.

3. Evaluation

a. Study area

This study is carried out for a region that covers the service domain of the U.S. NWS's Mid-Atlantic River Forecast Center (MARFC). Approximately, the region is bounded by latitudes 38° and 42°N and longitudes 70° and 83°W. As part of the upper East Coast of the United States, the region receives a bulk of the precipitation between the months of September and April from extratropical cyclones of synoptic scale (Maglaras et al. 1995), known as nor'easters. In the other months, mesoscale convective systems bring rainfall to the region and are a major cause of flooding. Occasionally, landfalling tropical storms can move into the region from the south and southwest.

b. Datasets

The forecast source is the precipitation reforecasts produced by running the GEFS version 9.0.1 retrospectively (Hamill et al. 2013). The GEFS issues 16-day-ahead ensemble forecasts. For the first 8 days, the reforecast model runs at a resolution of T254L42 on a quadratic Gaussian grid with a spacing of approximately 40 km. For forecast days 9–16, the reforecasts are saved at a resolution of T190L42 with a grid spacing of approximately 54 km. The reforecasts are saved on a 3-h time step for the first 72 h and a 6-h time step from 78 to 384 h. The reforecasts consist of one control and 10 perturbed members issued daily at 0000 UTC.

As verifying observations, we use an adjusted version of the NWS multisensor quantitative precipitation estimates (MQPE) available for 1997–2013. The original NWS MQPE dataset was created at the MARFC using the Stage III algorithm package for 1997–2001 and the Multisensor Precipitation Estimator (MPE) from 2002 onward (Seo et al. 2011; Zhang et al. 2011; Kitzmiller et al. 2011). The dataset contains hourly precipitation estimates on the Hydrologic Rainfall Analysis Project (HRAP; Greene and Hudlow 1982) grid, which is approximately 4.7 km in resolution in the midlatitudes. Interested readers are referred to Zhang et al. (2011) for a more comprehensive account of the process in creating this dataset. In the HRAP coordinate system, the study domain is rectangular in shape and has the boundary designated by points (850, 470) and (1046, 666). The spatial resolution for the observations is scaled up 4 times to about 19 km because of insufficient computational resources available for processing finer grids. As a result, the study domain comprises a grid of 50×50 cells. In this study, the hourly precipitation estimates are accumulated to 6-h values ending at synoptic hours (i.e., 0000, 0600, 1200, and 1800 UTC). A common rain gauge detection limit is 0.254 mm. Here, we use 0.25 mm as the threshold to distinguish between wet and dry conditions. Those aggregated values below this threshold are set to 0 in the postprocessing.

Substantial errors are found in the original MQPE dataset, as shown by several researchers (Zhang et al. 2011; Eldardiry et al. 2015). The most notable among these include a negative bias prior to 2003, related to a software error in the NEXRAD precipitation processing system, and the presence of anomalously large precipitation values stemming from erroneous gauge records. To address these inaccuracies, we use the gridded monthly gauge-based precipitation totals from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 2004) as the reference in the adjustment of the original MQPE dataset, following Zhang et al. (2011). As demonstrated in Zhang et al. (2011), the adjustment greatly improves the consistency between the Stage III algorithm and the MPE algorithm in bias characteristics. We consider the adjusted MQPE the best-quality precipitation product available for the validation effort here. Any remaining inconsistency would not have an effect on our conclusions for the postprocessed ensembles, since the adjusted MQPE is taken as ground truth and the model calibration as well as the SS schemes are performed against this common dataset.

c. Methods

The verification period is selected to be 1997–2013. For this time period, both the GEFS reforecasts and archived MQPE data are available. The rank structure of the SS-ENS is constrained by the number of ensemble members in the GEFS reforecasts, which is 11. For fair comparisons, 11-member ensembles are generated from the MMGD-SS procedure. These ensemble hindcasts are used in the evaluation of the three SS schemes. In addition, to evaluate the impact of the ensemble size on the spread of the processed ensembles, we also include 50-member ensembles generated from the MMGD-SS procedure using the SS-ANA scheme. To be consistent with the upscaled MQPE data in spatial resolution, the GEFS reforecasts are downscaled using nearest-neighbor interpolation for the edge grid cells and bilinear interpolation for the rest of the grid cells.

The MMGD model is calibrated using the GEFS reforecasts and MQPE datasets for the period of 1997–2013. Note that the SS schemes do not require calibration. Obtaining sufficiently large samples to reduce sampling uncertainty is of great importance in parameter estimation. To this end, a two-dimensional moving window of 61 days across the calibration years is used to pool forecast–observation pairs. The 61-day dimension is centered on the lead time of the reforecast. The choice of this window size, which also takes precipitation seasonal variation into consideration, is based on our past experience with calibrating the MEFP. A simple spatial data pooling scheme is also performed before parameter estimation. This is done by combining the samples of forecast–observation pairs from each of the grid cells in a 3×3 block. The pooled pairs are used to estimate parameters (at a 6-h time step) common to the grid cells in the block.

Hindcasting is performed for each 6-h time step in the period of 1997–2013. Spatially, hindcasting is run for each grid cell in the study area using the parameters of the 3×3 block the grid cell belongs to. We note here that in the hindcasting runs, the historical data for the current hindcast date is excluded for the SS-CLM and SS-ANA schemes to avoid giving an unfair advantage to the schemes in the validation. Specifically, the historical forecasts associated with the current hindcast date are excluded in the analog search for the SS-ANA scheme, and the historical forecasts from the date following the current hindcast date are used in the construction of the rank structures for the SS-CLM scheme.

Verification is performed for the period of 1997–2013. Thus, a dependent validation is conducted. The choice of this verification strategy is considered reasonable here for the following reasons: 1) As shown in Wu et al. (2011), results from a dependent validation and a cross validation of the MMGD model are very close, indicating that the MMGD model is reasonably robust. 2) The primary objective of this study is evaluating the

three SS schemes comparatively in terms of their spatial–temporal behaviors against a common setting for the MMGD model, rather than the overall sensitivity of the MMGD-SS modeling.

Ideally, the postprocessed ensembles can fully acquire the spatiotemporal properties of the natural weather regimes. In other words, each ensemble forecast field can be seen as a possible realization of the weather processes. If this is the case, then statistically, the postprocessed ensembles should be indistinguishable from the observations, not only at individual grid cells but also over the forecast domain. Our evaluation is based on this general postulate. Using correlation measures, we treat the observed fields and the fields of the ensemble forecasts as spatial or temporal stochastic processes and examine their autocorrelations. An ensemble member over the study domain will be called an ensemble field henceforth.

Since the SS schemes are rank based, the performance of these schemes is first assessed using the Spearman's rank correlation. This is done by examining how closely the Spearman's rank correlation coefficient of the processed ensembles tracks that of the observed data, spatially and temporally. The quality of the processed ensembles is further evaluated using a number of verification metrics. Among them, the Pearson's correlation is used to see how well this correlation measure can be preserved in the ensembles. In several previous studies, (Schaake et al. 2007; Wu et al. 2011; Brown et al. 2014a), the MMGD model has been shown to be capable of producing skillful precipitation ensemble forecasts. This capability is shown here using the BSS. The reliability diagram is a graphical method that can be used to reveal the reliability of a probabilistic forecast system (Jolliffe and Stephenson 2011, section 7.6.1). This tool is applied to the individual grid cells and areas of aggregated cells to assess the relative impacts of these SS schemes on the reliability of the postprocessed ensemble forecasts. The spatial patterns of the ensemble fields are assessed using the Baddeley's delta (Gilleland 2011). This is a binary image metric and can be used to measure similarity in spatial patterns between the ensemble fields and the corresponding observed fields. The metric is advocated as "ideal for users concerned with having very similar features (e.g., size, shape, orientation, and location)" in Gilleland (2011), where a comparison is conducted for the performance of nine 24-h forecasts of 60-min rainfall from various configurations of the Weather Research and Forecasting Model (a mesoscale numerical weather prediction system; https://www.mmm.ucar.edu/weather-research-and-forecasting-model). The Hausdorff distance is another spatial verification measure (Baddeley 1992; Schwedler and Baldwin 2011) we experimented with. In our evaluation, we find that the results obtained from these two methods are similar, and therefore only the results for Baddeley's delta are presented. Precipitation climatology varies spatially in a significant way in the study domain. The SS schemes are also compared for their capability of capturing this variability. Last, we show and examine a few ensemble fields against the corresponding observed field for a selected large precipitation event. This helps give the readers a sense of how the processed ensemble fields may behave.

The Baddeley's delta metric can be used to assess how well the forecast captures the observed spatial patterns. To use this metric to compare two precipitation fields with varying intensities at individual grid cells, the precipitation fields need to be converted to binary fields at multiple thresholds. Here, we give a simplified version of this metric (Gilleland 2011):

$$\Delta(A,B)=\left(\frac{1}{N}\sum\{w[d(\mathbf{x},A)]-w[d(\mathbf{x},B)]\}^2\right)^{1/2}, \qquad (1)$$

where A is a set of ones, B is another set of ones, and $\mathbf{x}$ is a grid point in the domain; $w(\cdot)=\min(\cdot, c)$, with c being a constant; $d(\mathbf{x}, A)$ is the shortest Euclidean distance between grid point $\mathbf{x}$ and set A; and the sum is taken over all N grid points in the domain. The $w(\cdot)$ is assumed to be eventually constant. Thus, the constant c in $w(\cdot)$ serves as a cutoff value, chosen in such a way that $w(t)=c$ for large t (Baddeley 1992).

4. Results

Evaluation results are obtained for the first three forecast days for which the GEFS is most skillful. The same verification metrics and computational settings are used in the computation for day 1, day 2, and day 3 forecasts. In this section, we present verification results mostly for day 1 forecasts, along with a selection of day 3 results, to give a fuller picture.

a. BSS for the individual grid cells

To show the skill of the postprocessed ensembles at the individual grid cells, we present the BSS of the ensembles at six precipitation exceedance thresholds. The reference used in the calculation is sample climatology. It is worth noting that at the individual grid cells, the evaluation outcomes only reflect the goodness of the MMGD modeling and the skill of the GEFS forecasts, regardless of what SS scheme is applied. Figure 1 depicts area averages of the BSS obtained at individual grid cells for day 1 ensemble forecasts against the corresponding observations at 6-h time steps, as a function of precipitation thresholds. The time window for the

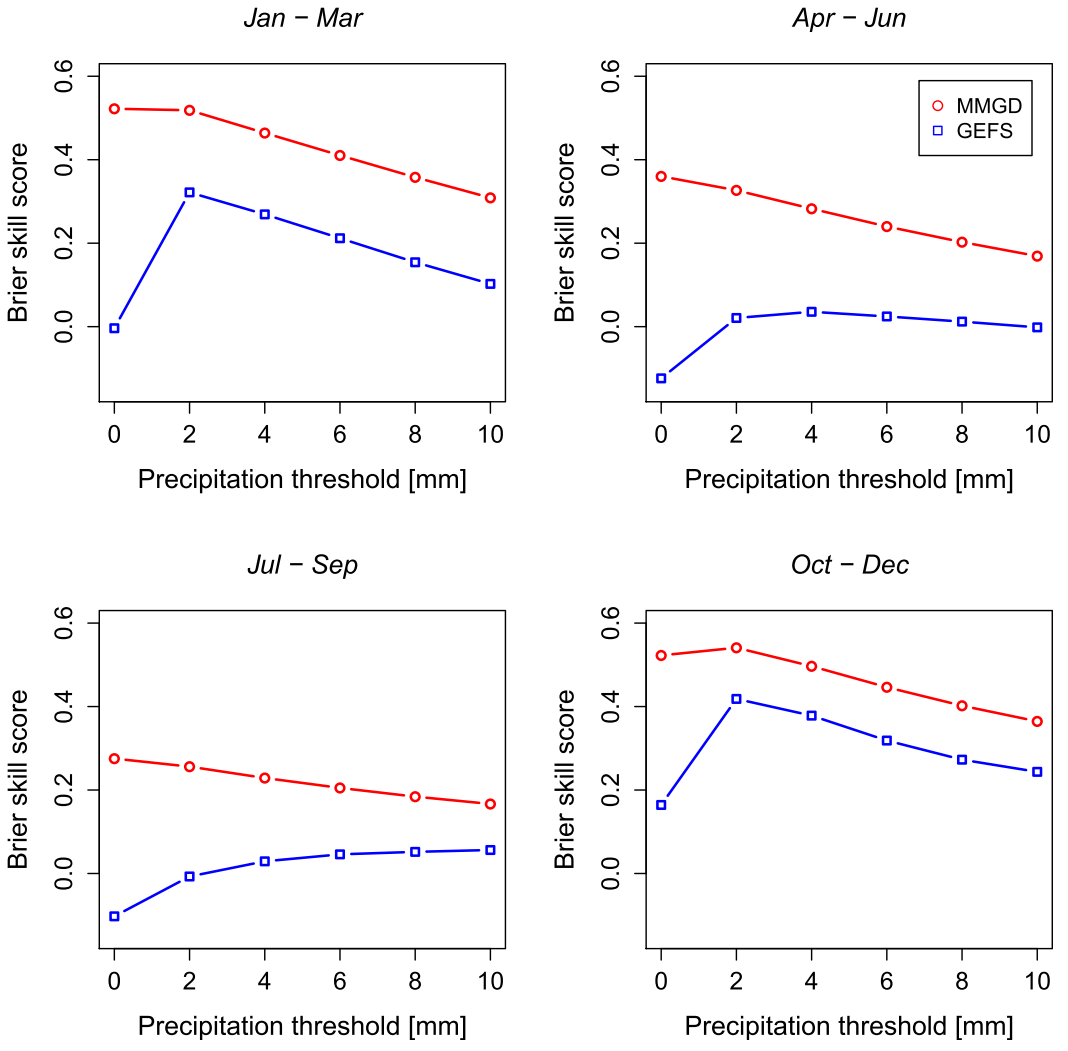

FIG. 1. Averages of the BSS values obtained at the individual grid cells over the central 30 × 30 area in the study domain for day 1 forecasts.

calculation is the years of 1997–2013 and the months indicated in the plots. As a result, the number of observations involved in the BSS calculation is about 6000. The averages are taken over the central rectangular region with end points at cell (11, 11) and cell (40, 40) of the domain. The BSS is bounded above by 1, with 1 being the perfect score. The figure shows that 1) the postprocessed ensembles are superior to the GEFS ensembles in this skill score, 2) these ensembles are more skillful for the cool season (October–March) than for the warm season (April–September), and 3) the BSS of the postprocessed ensembles decreases as the threshold increases. We note here that the GEFS values near 0 are nonverifiable because the degree of precision of the verifying MQPE values is at most as high as that of the rain gauges involved in producing the MQPE. A common rain gauge detection limit is 0.254 mm. In the postprocessing, we set 0.25 mm as the cutoff to distinguish between wet and dry conditions. Figure 1 shows that the MMGD postprocessing handles this cutoff well. By contrast, the abrupt behavior of the GEFS at the 0 threshold indicates that the GEFS is incapable of resolving dry–wet conditions at this cutoff.

b. *Spatial Spearman's correlation*

We then examine the spatial autocorrelation of the observed and ensemble fields, in terms of the Spearman's correlation, dependent on the distance between the grid cells (similar results are obtained for the Kendall rank correlation). Figure 2 shows area-averaged spatial correlograms for this measure for Day 1. The x axis represents separation between the cells. For

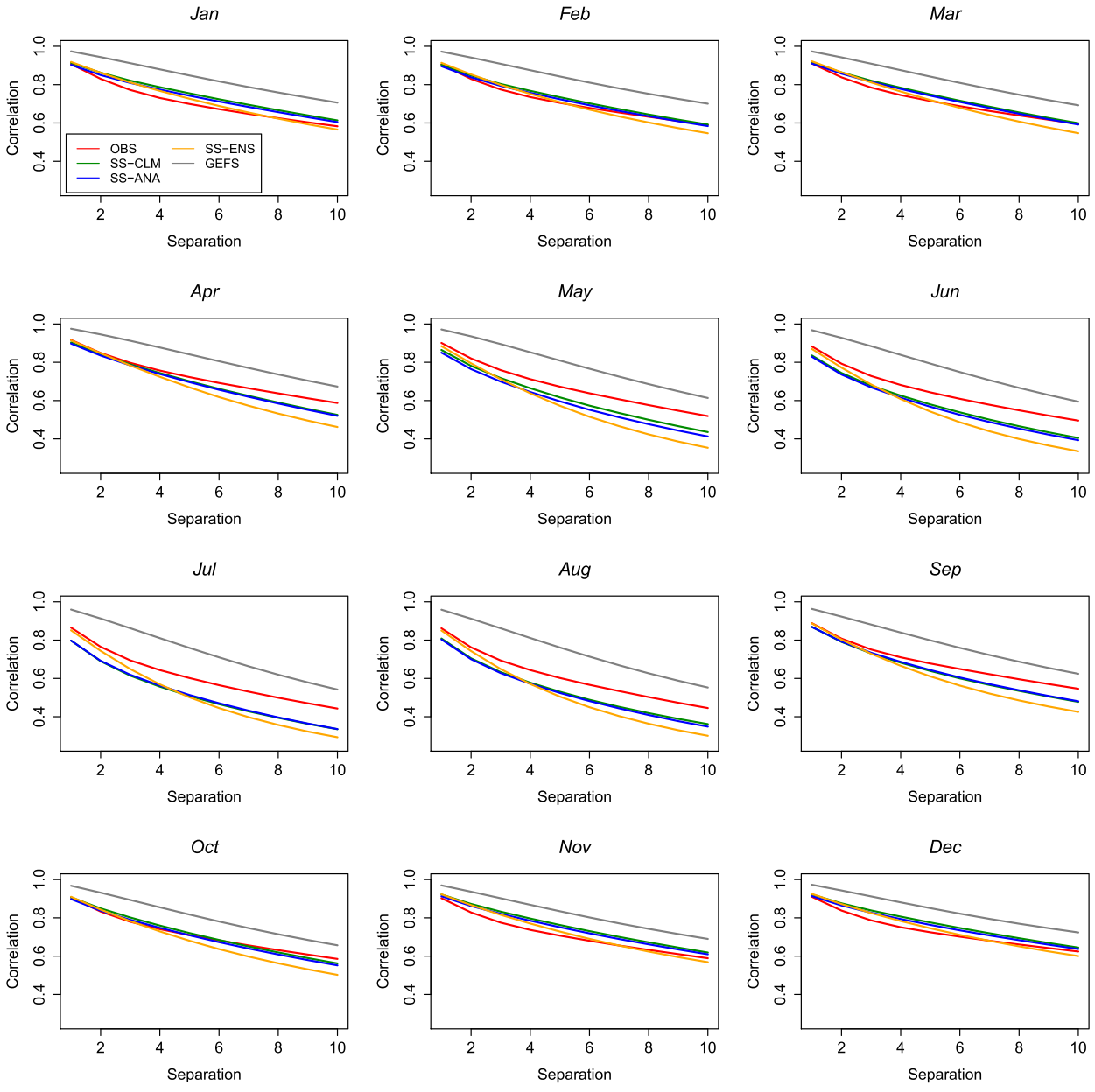

FIG. 2. Area-averaged spatial correlograms in terms of the Spearman's rank correlation between grid cells. The x axis represents separation between the grid cells. A separation of 1 signifies that the separated cells are immediately next to each other, etc. The results are obtained for day 1 forecasts over the 30×30 rectangular central area.

instance, if two cells are immediately next to each other horizontally or vertically, then they are said to have a separation of 1 (each cell is measured approximately 18.8 km in each dimension). Cells that are not horizontally or vertically aligned are excluded in the calculation. The y axis represents the correlation coefficient. The values in the figure are averages over a large central area to present an overall picture. To eliminate edge effects, the central area is chosen to be the rectangular portion of the study domain with end points at grid cell (11, 11) and grid cell (40, 40). To avoid multiple counting, for a given cell, only those cells to the right and below are used in the calculation. For a given cell, the sample is taken to be all the values from the days over the reforecast years (1997–2013) and a specific month. As a result, the sample size is about 2040 for the observed. Note that for the ensembles, only forecasts from 6 to 24 h are sampled. Moreover, for the ensembles, each member is

treated the same way as the observed, and then the resulting correlation values are averaged over the members. In the plots, the red line represents the observations; the gray line represents the GEFS ensembles; and the other three lines represent the postprocessed ensembles obtained with the use of the SS-CLM (green), SS-ANA (blue), and SS-ENS (orange). As expected, all the lines trend down as the distance increases. Noticeably, the spatial correlation of the GEFS ensembles is overly high, as evidenced by the wide margin between the GEFS line and the observation line, for each of the months. For the cool season of October–March, the lines of the postprocessed ensemble are in good agreement with the observation line. This coincides with the period when large-scale frontal systems occur most frequently in the region during the course of a year. For the warm season of April–September, the lines of the postprocessed ensembles deviate significantly from the observation line, exhibiting degradations in correlation to varying degrees as the distance increases. This behavior may be attributed to more frequent occurrences of random ranking in the SS (see section 2b for a discussion), resulting from lower predictability of smaller-scale systems. We observe that the SS-CLM line and the SS-ANA line stay close to each other. This would likely result from the fact that both the SS-CLM and SS-ANA use observations in constructing the rank structures.

c. *Spatial Pearson's correlation*

The Pearson's correlation provides a measure of linear correlation between two variables. Figure 3 shows area-averaged spatial autocorrelation values in terms of the Pearson's correlation for day 1. The four months in the plots are representative in Fig. 2 and selected here to facilitate comparisons. The results shown in this figure are obtained the same way as those shown in Fig. 2 except that the Pearson's correlation is used. For the postprocessed ensembles, we can see that the overall patterns are similar between the two figures. Overall, the SS-ENS tracks the observation the best among the three SS schemes for smaller gridcell distances (more conspicuous for day 2 and day 3; results not shown here). For the GEFS ensembles, unrealistically high correlations are, once again, present.

d. *Temporal Spearman's correlation*

Similarly, the Spearman's rank correlation is examined for temporal separation. Figure 4 shows averaged temporal autocorrelation for day 1 ensemble forecasts and the corresponding observations. The x axis represents separation between the 6-h time steps in day 1 of the forecasts. For instance, two time steps have a separation of 1 if they are immediately next to each other. The y axis represents the correlation coefficient. For a given grid cell, samples are taken over the reforecast period (1997–2013) and the months indicated in the plots. For a given separation number, the correlation values are area averages over the same rectangular area used for Fig. 2. Moreover, for the ensembles, correlation values are averaged over the 11 members. We can see that the temporal correlation is weak when the 6-h time steps are just a few hours apart, a reflection of the intermittent nature of precipitation events in the region. Overall, the SS-CLM and SS-ANA lines are closer to the observation line than the SS-ENS line. Note that the correlation values here are smaller than 0.5. When the correlation is low in the spatial cases, the SS-CLM and SS-ANA schemes also fare better. As with the spatial correlations, the temporal correlation values of the GEFS ensembles are also overly high.

e. *BSS for aggregated areas*

How would the postprocessed ensembles perform in terms of the BSS for aggregated areas? To answer this question, we examined four rectangular central areas of the study domain with aggregation sizes of 5×5, 10×10, 20×20, and 30×30. As an example, we show the case of 20×20 in Fig. 5. The x axis represents the thresholds. The y axis represents the score, with 95% confidence intervals included. Note that the error bars are shifted slightly away from the thresholds so that they can be easily seen. These confidence intervals are estimated via the percentile method of bootstrapping (Chernick 2008). For each of the ensemble fields and the corresponding observed field, the mean value is calculated over all grid cells in the 20×20 area for any given 6-h time step in day 1 of the forecasts. The BSS is calculated with these mean values in the time window of the GEFS reforecast years (1997–2013) and the months indicated in the plot, with a sample size of about 6000. Sample climatology is used as the reference forecast. We can see that the results of the three SS schemes are very close. This verification measure is not sufficiently sensitive to misspecification of spatial correlations to reveal a clear performance ranking of the different SS schemes. The postprocessed ensembles outperform the GEFS ensembles for the lower thresholds. However, the postprocessing tends to become less effective toward the higher thresholds. Note that the postprocessed ensembles always have shorter error bars than the GEFS ensembles. Similar behaviors are observed for the other aggregation levels (results not shown). The GEFS results are more uncertain, partly due to the smaller effective sample sizes used in the computation because of the GEFS downscaling, and partly due to a lack of calibration.

f. *Reliability for aggregated areas*

Reliability is a key criterion for assessing the quality of ensemble forecasts (Jolliffe and Stephenson 2011).

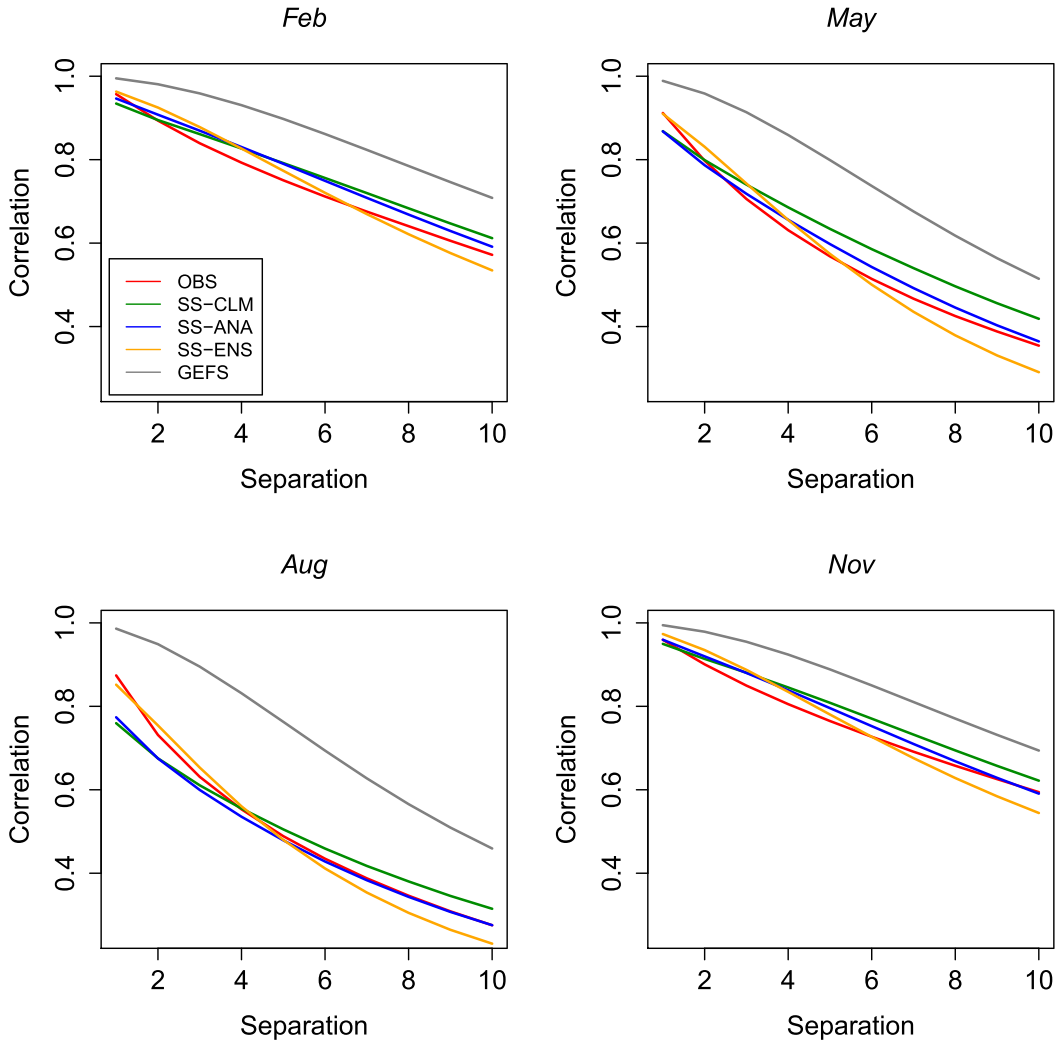

FIG. 3. Area-averaged spatial autocorrelation in terms of the Pearson's correlation. The results are obtained the same way as those shown in Fig. 2 except that the Pearson's correlation is used.

Several previous studies on the MMGD model (Wu et al. 2011; Brown et al. 2014a) show that the model is capable of producing reasonably reliable precipitation ensemble forecasts at single forecast points. This is also true for the grid cells we have examined in this study. Note that at the individual grid cells, rearranging the ensemble members does not change the sample used for computing skill metrics. A question that is important to ask is whether the level of reliability achieved by the postprocessed ensembles at the individual grid cells can be preserved for larger areas consisting of multiple grid cells. The question may be addressed by the following case. Figure 6 shows reliability diagrams for four central areas of various sizes in the study domain. The threshold value used corresponds to the 95th percentile of the observed data, around 4 mm depending on the aggregation level. The four selected areas are the grid cell at (25, 25), a central rectangular area consisting of 10 grid cells in both directions (10×10), the central 20×20 area, and the central 30×30 area. For each of the ensemble fields and the corresponding observation field, the mean value is calculated over all grid cells in the central areas for every 6-h time step in day 1 of the forecasts. The reliability diagrams are generated with these mean values in the time window of the GEFS reforecast years (1997–2013) and the months of October–March, with a sample size about 12 000. Along with the reliability curves, 95% confidence intervals are shown at the five probability bins. These confidence intervals are estimated via the percentile method of bootstrapping (Chernick 2008). The confidence intervals at the lowest forecast probabilities are too small to be visually

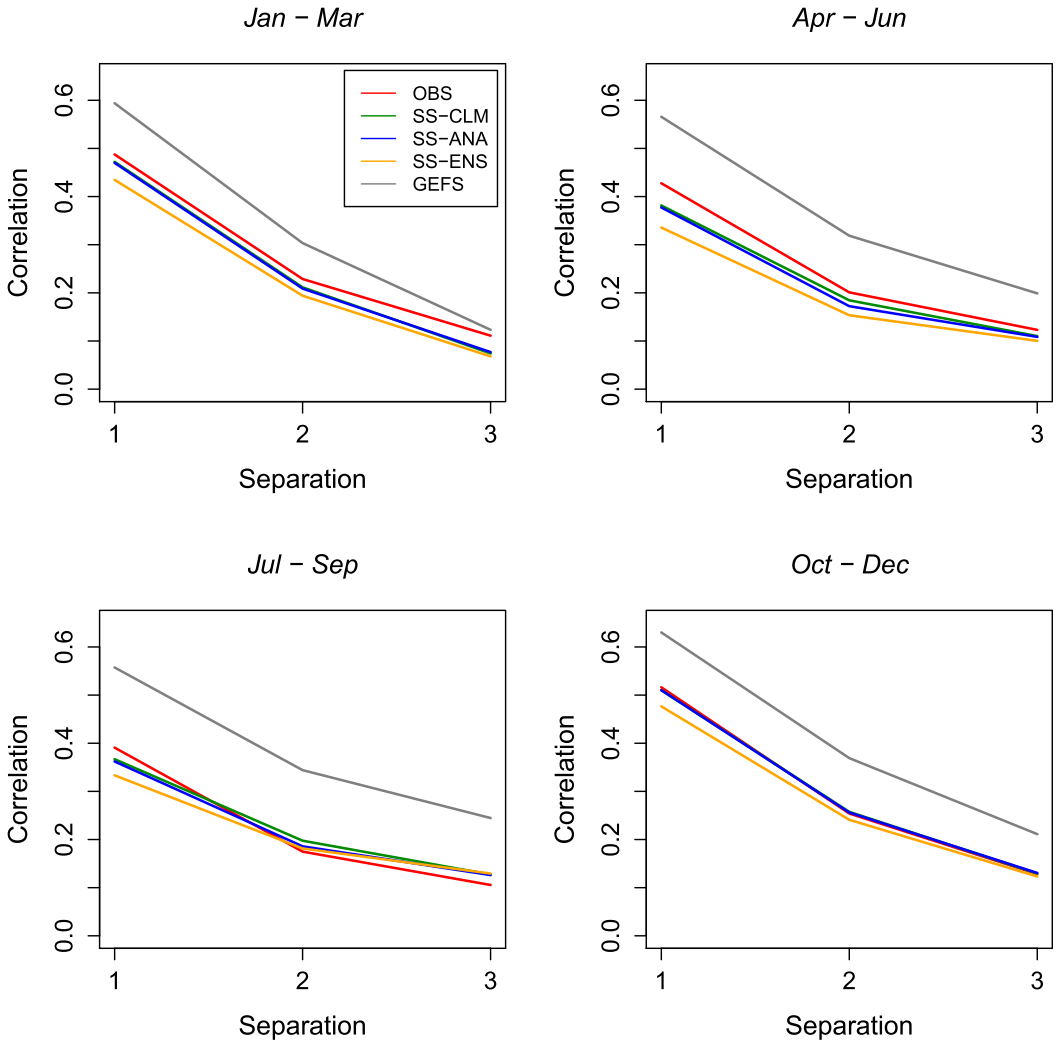

FIG. 4. Area-averaged temporal autocorrelation in terms of the Spearman's rank correlation between the 6-h time steps in day 1 of the forecasts. Results are area averages over the same rectangular area used for Fig. 2.

discernible. Note that the SS-CLM (green) at the highest forecasted probability in the 30 × 30 panel appears without an accompanying confidence interval. This happens because at the threshold for this reliability diagram, the binary observed values are all 1, resulting in a 0-width confidence interval at this probability. We can see from the figure that the GEFS ensembles are quite poor at higher forecast probabilities. There is only a single line in the upper-left panel for the postprocessed ensembles because none of the SS schemes has any effect on the individual grid cells in terms of computing skill metrics. It can be seen that the postprocessed ensembles tend to become less reliable as the aggregation level increases. The degradation results in underforecasting of larger probabilities. This tendency is also seen in other cases with different seasons and thresholds. The degradation is unlikely a consequence of the small ensemble size being used, as the same behavior can be seen as well for a size of 50 in the SS-ANA case. It may stem from the random ranking discussed in section 2b. Random ranking can occur for precipitation in applying the SS schemes due to the presence of many dry events. In applying these schemes, some of the large ensemble members at the individual grid cells can be randomly placed in certain ensemble fields, resulting in reduced mean values for some ensemble fields. Further study is needed to gain better understanding of this problem.

g. *Spatial similarity*

While a good tool for assessing how well an individual forecast field captures spatial patterns of the corresponding observed field, the Baddeley's delta metric is

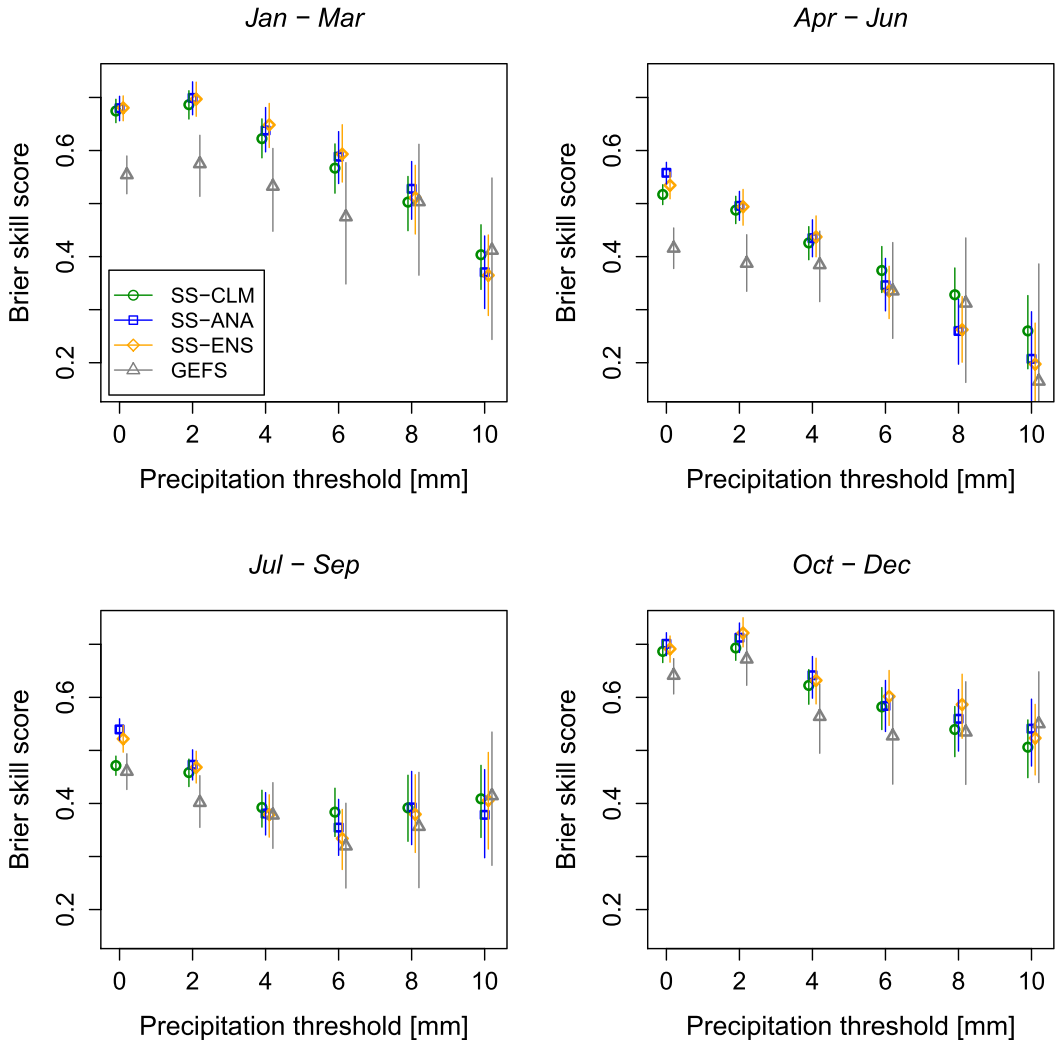

FIG. 5. BSS (with 95% confidence intervals) for forecast fields over the central 20 × 20 area in the study domain. The results are obtained for day 1 ensemble forecasts. They are shifted slightly away from the thresholds for clarity.

also useful for summarizing the overall spatial similarity performance of forecasts when aggregated over many cases (Gilleland 2011). Figure 7 shows averaged Baddeley's delta between the observed field and the corresponding ensemble fields computed using all of the data. The *x* axis represents the thresholds used in the computation. The observed and the ensemble fields are turned into binary fields with respect to a given threshold. The *y* axis represents the metric with smaller values indicating greater similarity. The metric is computed using the locmeasures2d function in the R (R Core Team 2016) package SpatialVx (version 0.4) from the Comprehensive R Archive Network (CRAN) repository (https://cran.r-project.org). The metric is first computed for each pair of an observed field and an associated ensemble field. Then a mean Baddeley's delta is calculated over the 11 ensemble members, the monthly time window for all of the reforecast years, and the months indicated in the plots. Here, the results for day 1 ensemble forecasts are displayed. For the extended spatial scales and the months of July–September, the SS-ENS outperforms the SS-ANA, which in turn outperforms the SS-CLM, for the lower thresholds. The performance difference narrows and vanishes as the threshold increases. Note that the GEFS ensembles outperform the postprocessed ensembles, which seems counterintuitive. However, the fact is that this verification scenario includes small-scale precipitation systems. The effectiveness of the SS schemes is reduced when dealing with these small-scale systems, due to their low predictability, which can lead to more frequent occurrences of random ranking in the application of these

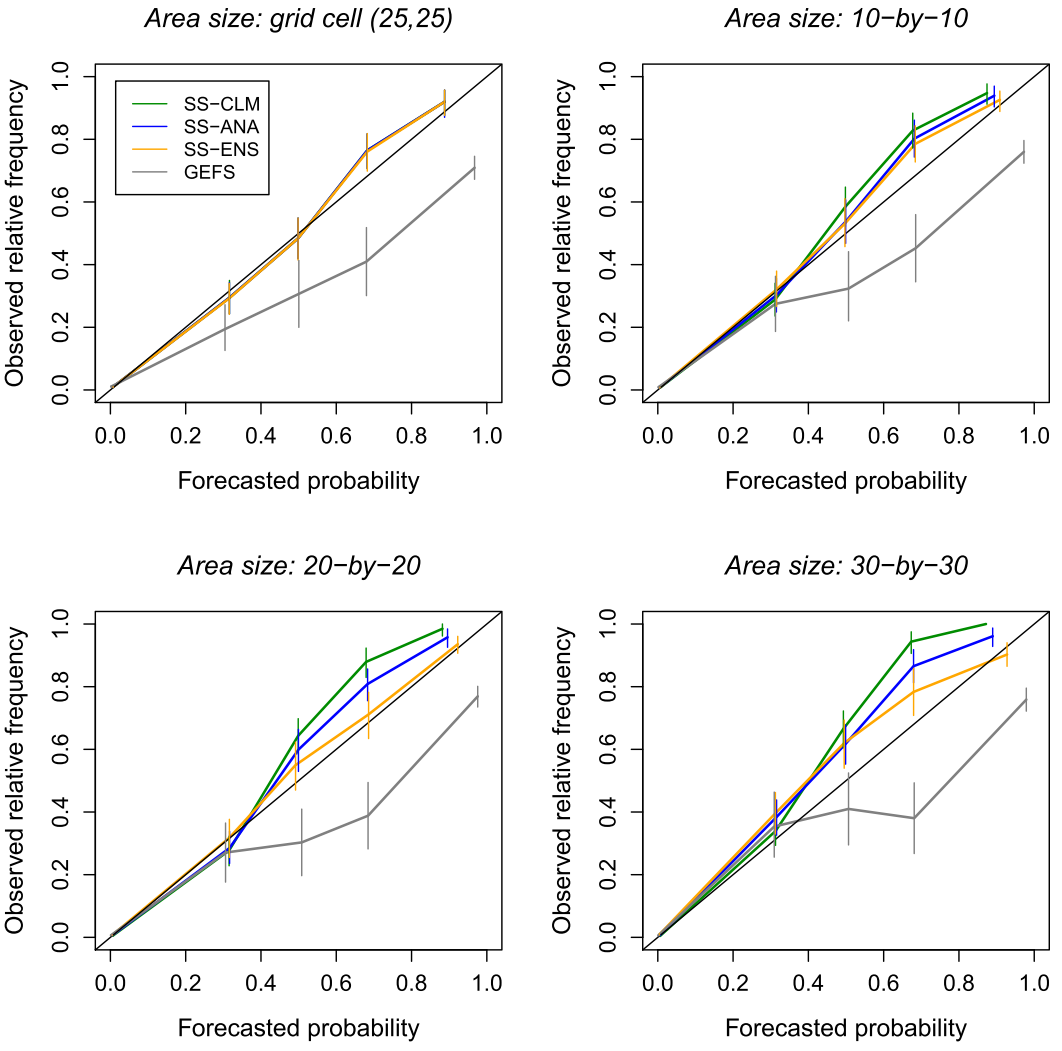

FIG. 6. Reliability diagrams (with 95% confidence intervals) for four central areas of various sizes in the study domain. The threshold value used corresponds to the 95th percentile of the observed data, around 4 mm depending on the aggregation level. The results are obtained for day 1 ensemble forecasts over the cool season.

schemes. Further investigation is made for the scenario of large precipitation events: those observed fields falling in the upper 10% of the regionally summed 6-h precipitation amounts (results not shown). The results also show clearly an underperformance of the SS-CLM scheme relative to the other two SS schemes for the two cases of extended spatial scales for the summer months at lower thresholds. Interestingly, for these two cases, the SS-ANA and SS-ENS schemes outperform the GEFS considerably.

h. Spatial variability

Precipitation climatology varies considerably in the study area. As an example, Fig. 8 shows temporally averaged 6-h observed precipitation amounts, obtained for the time steps in August across the years of 1997–2013 at the individual grid cells over the study domain. Evidently, a vertical swath of area with much higher average precipitation runs across the domain. How would the three SS schemes differ in capturing this spatial variation? Here, we examine the closeness between the climatology of the observed and that of the ensemble forecasts over the study domain using the RMSE. The results are shown in Fig. 9 for day 1. The RMSE is computed for each month, as indicated by the x axis of the figure. The y axis gives the mean RMSE with 95% confidence intervals included. The mean RMSE is obtained as follows. For the years of 1997–2013, the 6-h observed precipitation amounts in each month are averaged at each of the grid cells in the study domain. We thus obtain a mean field of observed precipitation. The same is performed for each of the ensemble members

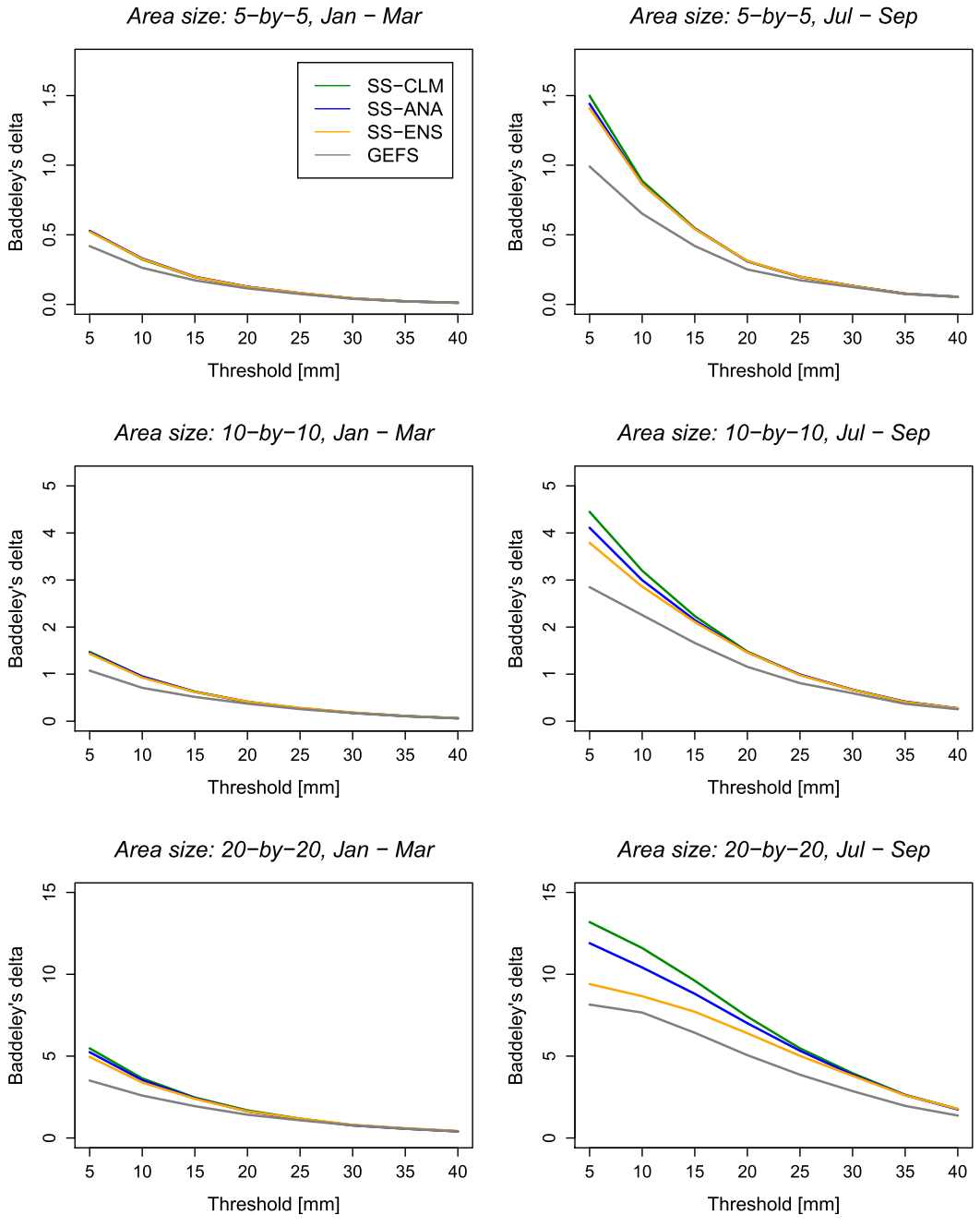

FIG. 7. Mean Baddeley's delta for a central (top) 5×5, (middle) 10×10, and (bottom) 20×20 area in the study domain for day 1 ensemble forecasts. The metric is measured in number of grid cells, with each cell measuring approximately 18.8 km in each dimension.

from a given SS scheme or the GEFS for day 1. Then, the RMSE is computed between the mean fields of the observed and the ensemble members over the study domain. Last, the RMSEs are averaged over the 11 ensemble members. The confidence intervals are estimated via the percentile method of bootstrapping (Chernick 2008). We can see from the figure that the SS-CLM scheme has much larger mean RMSEs than the SS-ANA and SS-ENS schemes throughout the months, indicating that the SS-CLM is less capable of correctly distributing precipitation in space. We also see that for the months of June–October, the GEFS ensembles perform better than the postprocessed ensembles.

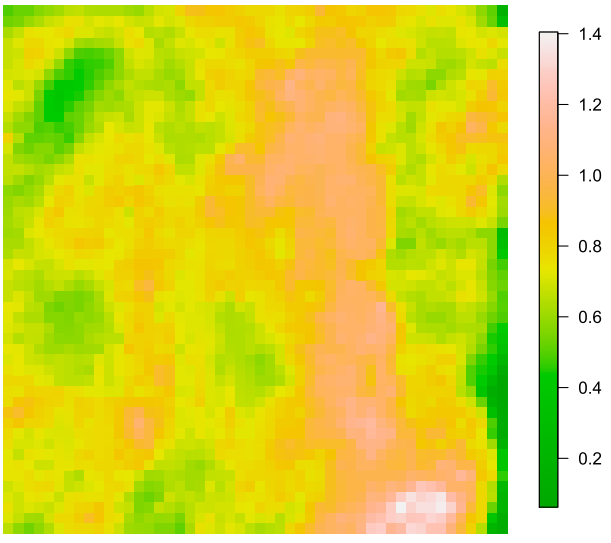

FIG. 8. Temporally averaged 6-h observed precipitation amounts (mm) for August over the study domain. The image is shown on the HRAP grid with the lower left corner at (850, 470).

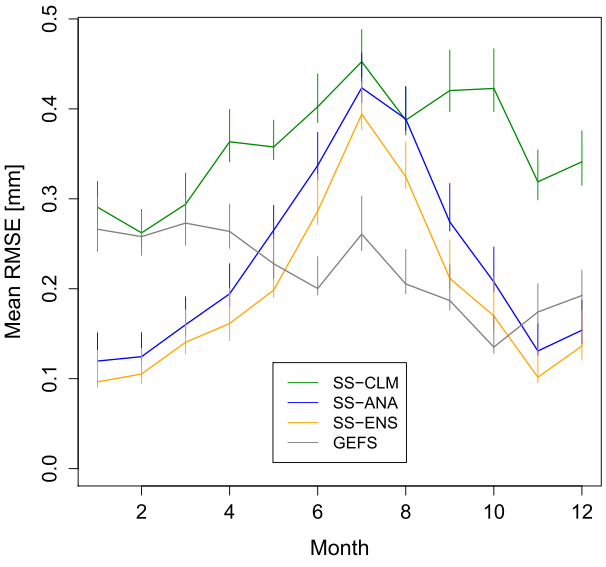

FIG. 9. Mean RMSE (with 95% confidence intervals) between the 6-h precipitation climatology of the observed fields and that of the ensemble forecast fields of the study domain.

i. Mean CRPS for aggregated areas

The continuous ranked probability score (CRPS) is one of the most widely used performance measures in ensemble forecast verification. The mean CRPS measures the overall quality of probabilistic forecasts as the expected squared error of the forecast probabilities for all possible events (Jolliffe and Stephenson 2011). Figure 10 shows the mean CRPS for the postprocessed ensemble fields against the observed fields at two aggregation levels. The CRPS is computed using the method described in Hersbach (2000). Data samples used in the computation are obtained the same way as those for Fig. 5. The *x* axis represents the lead times. The *y* axis represents the metric with smaller values indicating better performance. As expected, the performance of the ensembles decreases with increased lead times. Also, the postprocessed ensembles outperform the GEFS ensembles in all of the cases. Among the postprocessed ensembles, differences are small or hardly discernible, indicating that this measure does not respond well to spatial-structure changes of the ensemble fields.

j. Verification results for longer lead times

Having shown verification results mostly for day 1 forecasts, here we extend our analysis to longer lead times, specifically, day 2 and day 3 forecasts. The same verification metrics and computational settings used for day 1 are employed for day 2 and day 3. Our general observation for the newer results, in comparison with the day 1 results, is that the performance of the SS schemes declines steadily as the lead time increases. A selection of results from day 3 is shown in Fig. 11 to give a cross-section view. In the verification results presented above, the abbreviation SS-ANA might be as well denoted as SS-ANA-D1 to signify that the scheme operates on day 1 forecasts in search of matching analogs. Here, results of a variant, SS-ANA-D3, are also included in the figure. The SS-ANA-D3 operates on day 3 forecasts to search matching analogs. As a result, the starting dates for the shuffling selected by the SS-ANA can differ from those selected by the SS-ANA-D3. The objective of including the SS-ANA-D3 is to see how much gain a lead-time-specific analog search scheme may have over the SS-ANA for that specific lead time. Figure 11a shows area-averaged spatial Pearson's correlation for February, corresponding to the panel of "Feb" in Fig. 3 for day 1. In comparing the two panels, we can see that the SS-CLM and SS-ANA lines for day 3 exhibit larger deviations from the observations (OBS) line at smaller separation numbers. This is typical for day 2 and day 3. Figure 11b shows values of the area-averaged temporal Spearman's rank correlation, corresponding to the panel of "Jan–Mar" in Fig. 4 for day 1. A larger deviation from the OBS line is seen for the SS-ENS scheme for day 3. Figure 11c shows BSS for forecast fields, corresponding to the panel of "Jan–Mar" in Fig. 5 for day 1. This plot shows only the BSS values without giving confidence intervals. Figure 11d shows a reliability diagram at the threshold of the 95th percentile of the observed data, corresponding to the panel of "Area size: 30 × 30" in Fig. 6 for day 1. This panel, along with other reliability results for day 2 and day 3 (not shown), provides additional evidence that the postprocessed

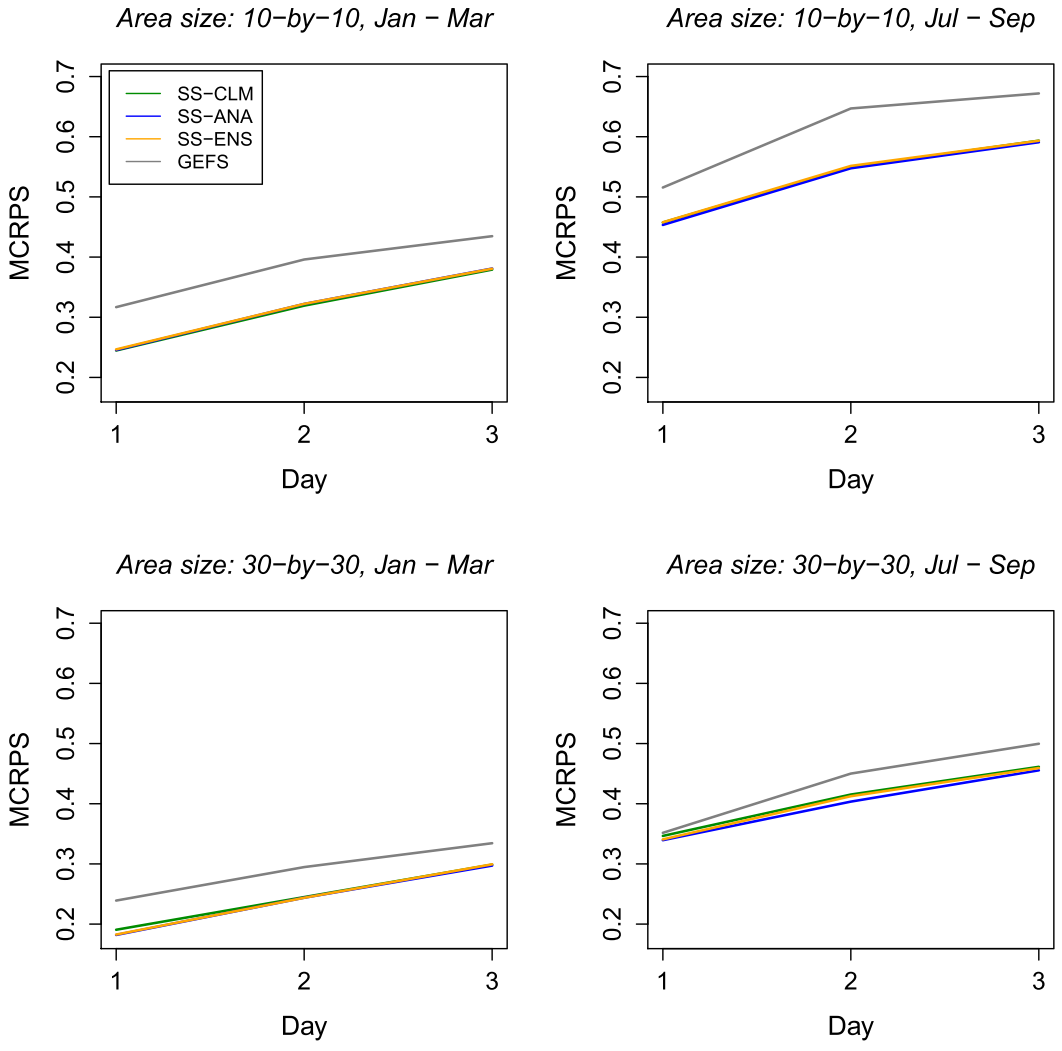

FIG. 10. Mean CRPS for aggregated areas for the lead times from day 1 through day 3.

ensembles tend to become less reliable as the aggregation level increases, an observation made for day 1. Figure 11e corresponds to the panel of "Area size: 20 × 20, Jul–Sep" in Fig. 7 for day 1. By comparing these two, we see that this metric for day 3 is increased for the lower thresholds as a result of an increased lead time. Another observation from this panel is that the SS-ENS outperforms the SS-CLM and SS-ANA. Figure 11f shows spatial variation RMSE, corresponding to Fig. 9 for day 1. Apparently, the overall relative performance between SS-CLM, SS-ANA, and SS-ENS is maintained for day 3 with the SS-ENS being consistently better than the SS-ANA, which in turn is consistently better than the SS-CLM.

k. A large precipitation event

Last, for a selected large-scale observed event, we examine whether the ensembles from various sources discussed above capture the observed event in terms of area-averaged precipitation amounts and gridcell maximum precipitation amounts. The observed event is the 6-h accumulation ending at 2400 UTC 16 April 2011 over the study domain. Its area-averaged amount is 12.01 mm, with heavy precipitation occurring at a number of grid cells. The largest gridcell amount is 76.41 mm. Corresponding to this event, Table 2 presents results of the GEFS ensembles, the processed ensembles obtained with the SS-CLM, SS-ANA, and SS-ENS schemes, each generated with 11 members, and the SS-ANA scheme with 50 members. For the ensemble sources, columns 2–4 of the table show area-averaged precipitation amounts for the largest ensemble member, the second-largest member, and the smallest member, respectively. The last column shows gridcell maximum amounts among all members over the study domain. Field images of the

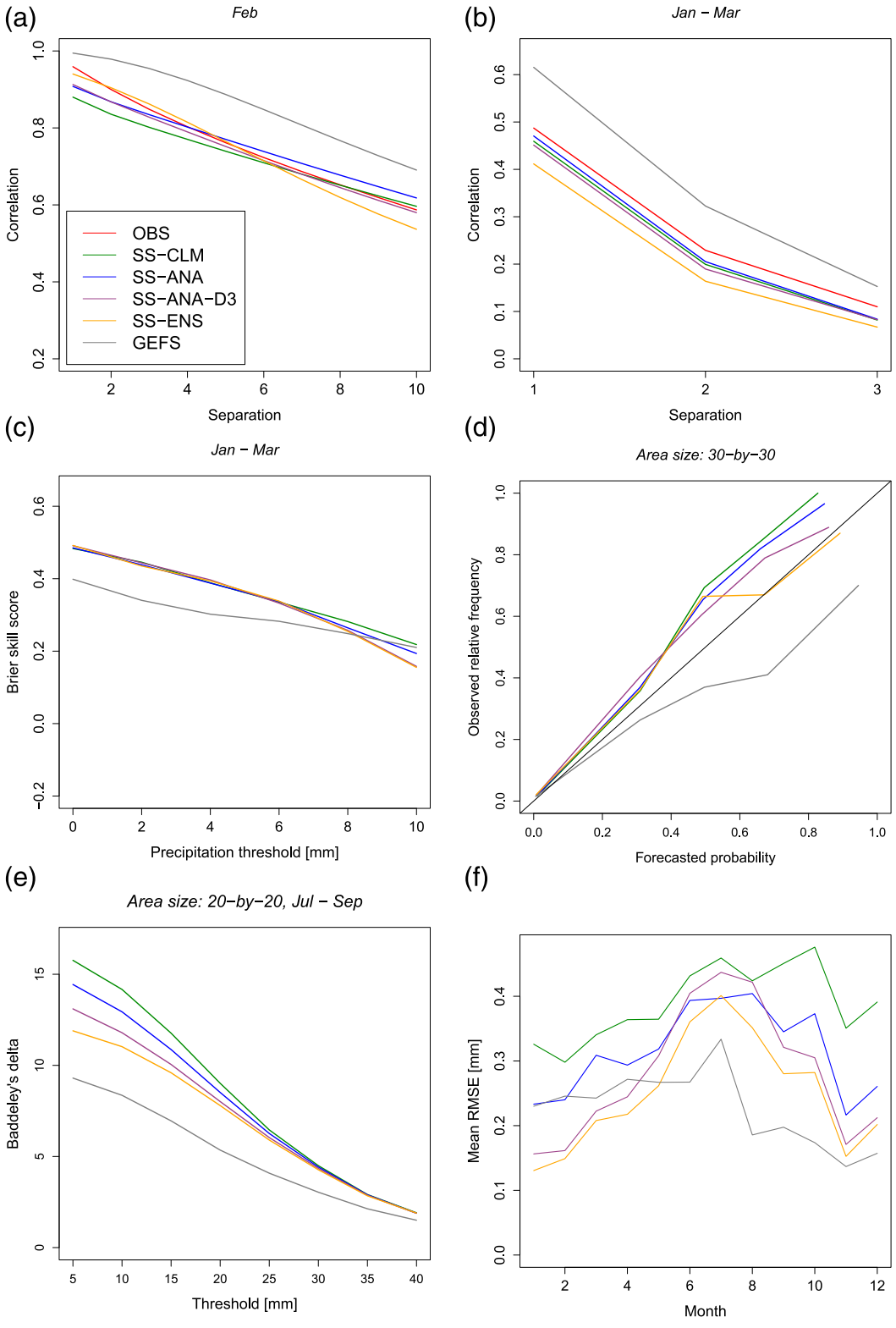

FIG. 11. A selection of verification results for day 3.

TABLE 2. Area-averaged precipitation amounts (mm) for the first two largest ensemble members and the smallest member, with gridcell maximum amounts among all members (mm).

Source	Avg (largest)	Avg (2nd largest)	Avg (smallest)	Max
OBS	12.01			76.41
GEFS	10.35	10.10	8.44	39.83
SS-CLM	9.15	8.16	1.82	38.31
SS-ANA	6.75	6.21	1.49	38.31
SS-ENS	7.04	5.87	3.21	38.31
SS-ANA (50 members)	10.23	9.17	1.75	54.89

observed event and the corresponding largest ensemble member of the various sources are also displayed in Fig. 12. We can make the following observations from the results:

1) The GEFS underestimates the observed area average by a large amount, at about 13% for its largest ensemble member. Its maximum gridcell value is much smaller than that of the observed. Its ensemble spread is small relative to the other sources.
2) The SS-CLM also underestimates the observed area average. Its gridcell maximum, being identical to that of the SS-ENS and SS-ANA, is only about half of that of the observed.
3) The SS-ANA and SS-ENS area averages are much smaller than the observed area average.
4) With the ensemble size increased to 50, the average and maximum values of the SS-ANA are also significantly increased, although still short of capturing the observed values. The important role of the ensemble size cannot be overemphasized since it has an impact on the prediction of large-scale events with heavy precipitation amounts, as demonstrated by this example.

In generating the postprocessed ensembles, a stratified sampling scheme is used to generate probabilities in the interval of (0, 1) [see Wu et al. 2011, Eq. (A7)]. The scheme specifies a probability for the ith positive ensemble member as $i/(n+1)$, where n is the number of positive ensemble members in an ensemble. This scheme corresponds to the commonly used Weibull plotting position for a normal Q–Q plot (Sonia et al. 2012). An issue with stratified sampling schemes in general is that the probabilities generated are bounded above by a number depending on n, in this case by $n/(n+1)$. For example, with $n = 11$, the largest probability value attained by this scheme is about 0.92, leaving possible larger values unsampled. This shortcoming may be addressed by selecting a suitable ensemble size during model calibration that takes historical observations and physical constraints into consideration. However, we also recognize that calibrating this parameter can be difficult since ensemble spreads are related to rare-event forecasting, which is extremely challenging to model. For the SS-ENS scheme, obtaining a large ensemble size is not as straightforward as it would be with the other two SS schemes because of the constraint imposed by the ensemble size of dynamic models. To overcome this potential issue, one will probably have to gather ensemble forecasts from several sources.

5. Discussion

Our primary reason for choosing the MMGD model in this study is that the model performed reasonably well in several previous studies (Wu et al. 2011; Brown et al. 2014a). The MMGD-SS approach fits into the empirical copula framework (section 2a), in which the MMGD model can be replaced by any suitable univariate postprocessing model. It would be interesting to compare how different univariate models perform in the empirical copula framework with different SS schemes applied.

A frequent concern arising from operational forecasting is that numerical weather prediction models can undergo major upgrades every few years, which potentially prevents them from providing adequately long and consistent historical datasets for postprocessing. A relevant question here is how the length (number of years) of historical datasets affects the performance of the SS schemes. Since the SS is part of the MMGD-SS postprocessing procedure, the problem has to be investigated in the MMGD-SS setting. A sensitivity study was carried out in Brown (2015) for the case of coupling the MMGD with the SS-CLM scheme for a basin-based postprocessing system. In the study, about 50 years of historical observations were available for performing the SS-CLM. The findings of the study include the following: 1) for precipitation, reasonable estimates of the MMGD parameters can be obtained from as little as 5 years of data, and 2) the sensitivities of the GEFS-based postprocessed precipitation forecasts to the length of calibration data are relatively small, both for the dependent and independent validation scenarios across a broad range of precipitation thresholds. In a

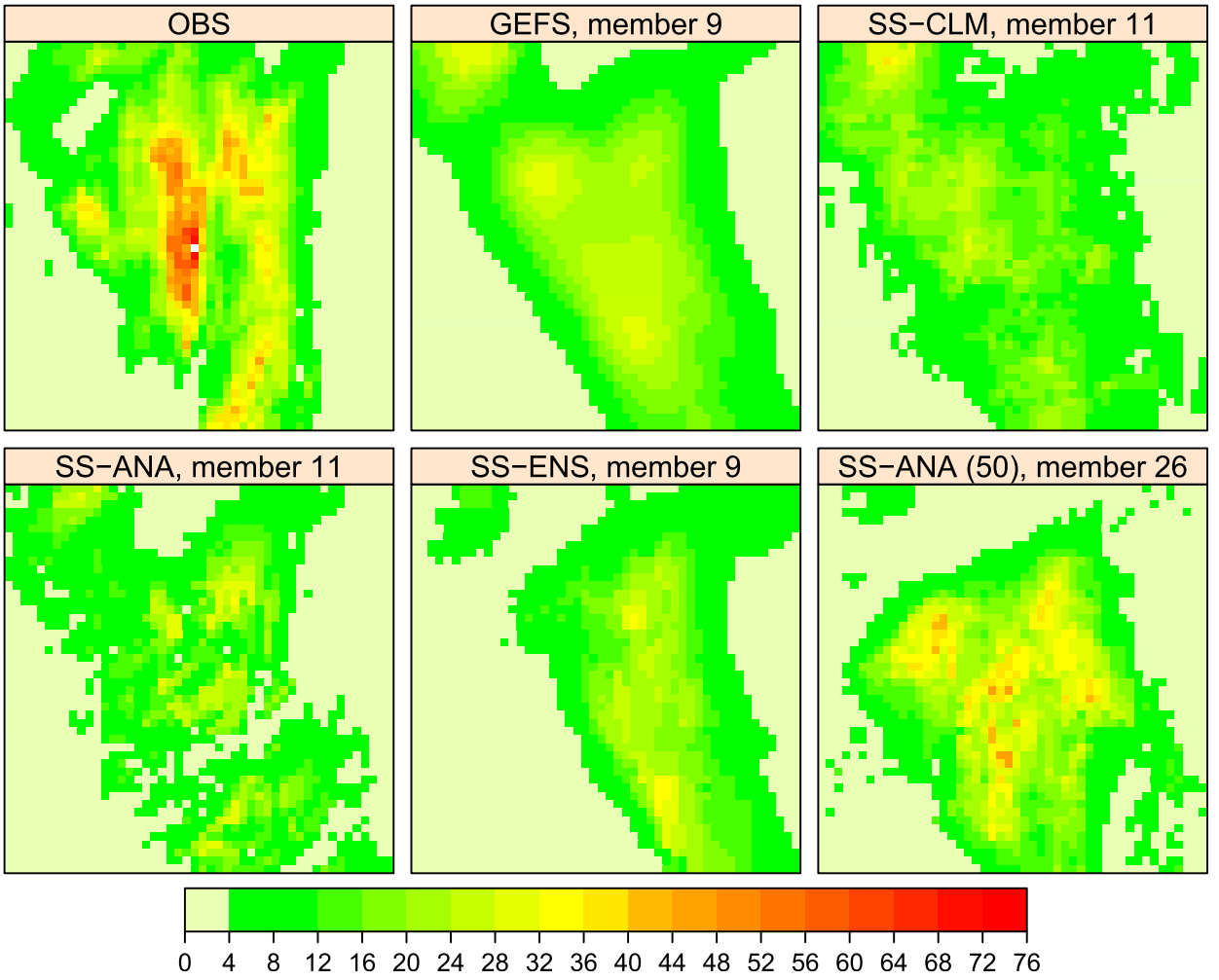

FIG. 12. Comparison of a large-scale observed event and the corresponding largest ensemble members from different sources. The observed event is the accumulation (mm) during 1800–2400 UTC 16 Apr 2011 over the study domain. The images are shown on the HRAP grid with the lower left corner at (850, 470).

gridded-forecast setting, however, we have not seen similar studies performed for the MMGD-SS. Nonetheless, the study provides a valuable reference for the MMGD-SS sensitivities.

A number of questions raised by this study warrant future research. Perhaps the most prominent one is the following: Why do the SS schemes tend to lose reliability with increased aggregation levels, as indicated by the results of the reliability diagrams? The practice of aggregation by pooling forecast–observation pairs from individual grid cells to assess postprocessing models tends to yield better verification results. However, since our objective is to assess ensemble forecast fields rather than ensemble forecasts at individual grid cells, we performed the aggregation differently by averaging forecasts and observations from individual grid cells for each of the postprocessed ensemble fields before the computation for the reliability diagram. The tendency to lose reliability suggests a limitation of these SS schemes. Two pertinent questions are 1) is the decay in reliability a result of the random ranking in the SS schemes, and 2) is postprocessing at individual grid cells in an area equivalent to, or consistent with, postprocessing at the area treated as one larger grid cell?

This study is limited in scope to short-range forecasts in a particular climate region at a specific spatial scale and temporal scale. In addition, a dependent validation strategy is used. These limitations will be addressed in future studies.

6. Summary and concluding remarks

The Schaake Shuffle method and its variants offer a practical approach for preserving spatiotemporal

coherence of multiple weather variables. Fundamentally, these methods exploit the Spearman's rank correlation between the weather variables of interest in a multivariate setting, such that the postprocessed ensembles can acquire the spatiotemporal properties of the observed climatology and, as individual forecasts, reflect the current atmospheric conditions in spatial gradient and temporal persistence. In this study, the Schaake Shuffle method and its two variants (SS-CLM, SS-ANA, and SS-ENS) are evaluated and compared in postprocessing GEFS precipitation ensembles. The MMGD model together with an SS scheme offers an empirical copula for modeling the relationship among multiple variables, in which the SS scheme provides the dependence structure. Moreover, since precipitation is intermittent in nature, the rankings produced by these SS schemes may not be unique. This can lead to inconsistent spatiotemporal variability and correlations in the postprocessed ensembles.

This study is carried out using data for the MARFC region. An adjusted version of the NWS MQPE is used as verifying observations. The forecast source is the GEFS precipitation reforecasts (version 9.0.1). The common period (1997–2013) of the MQPE records and the GEFS reforecasts is used. A summary of the findings on the performance of the MMGD-SS with the three SS schemes is given below.

1) For the months of November–March, the averaged spatial autocorrelation, in terms of the Spearman's rank correlation, is in good agreement with that of the observed fields, for all of the gridcell distances considered. For the other months, the agreement deteriorates as the distance increases. This performance change corresponds well to the change of the dominant precipitation systems from the synoptic type in the cool season to the mesoscale type in the warm season, which suggests that the performance of the SS schemes depends on the scale of precipitation systems.
2) The SS-ENS tracks the observation better than the SS-CLM and SS-ANA in terms of the spatial Pearson's correlation for shorter grid-cell distances.
3) The averaged temporal autocorrelation, as measured by the Spearman's correlation, declines sharply with temporal separation and is weak when the separation is just a few hours. The SS-CLM and SS-ANA schemes perform better than the SS-ENS in terms of this correlation.
4) The BSS (a measure of overall skill) for the processed ensembles over an aggregated area increases as the level of aggregation increases.
5) Despite the gain in the BSS stated above, reliability of the processed ensembles over an aggregated area tends to degrade as the level of aggregation increases. This degradation results in underforecast probabilities.
6) For large spatial scales and the summer months, the SS-ANA and SS-ENS outperform the SS-CLM, in terms of the Baddeley's delta.
7) Precipitation climatology varies spatially in the study area. The SS-ENS scheme outperforms the SS-ANA, which in turn outperforms the SS-CLM, in capturing this spatial variability in terms of the RMSE.
8) A proper choice of an ensemble size in the MMGD-SS approach is of great importance as it affects the spread of the postprocessed ensembles. If we treat the ensemble size as a parameter, then it may be determined through some calibration process that takes historical observations and physical constraints into account.
9) As the baseline of our evaluation, the GEFS ensembles underperform in most verification measures considered in this study.
10) The presence of tied ranks in the rank structures compromises the effectiveness of the SS schemes.

Apparently, some verification metrics are more capable of differentiating the three SS schemes than others. The averaged spatial climatology RMSE shown in Fig. 9 reveals clear differences among the three SS schemes. Although to a lesser degree, significant differences can also be seen for the results in Fig. 6 (reliability diagram) and Fig. 7 (Baddeley's delta) for the large aggregation levels and the warm season. On the other hand, differences are marginal in Fig. 5 (BSS) and Fig. 10 (mean CRPS). These two metrics are apparently much less sensitive to the changes in spatial structures of the ensemble fields.

A straightforward ranking of the three SS schemes does not seem attainable. These schemes give mixed results by the correlation measures. For instance, the SS-ENS performs better than the other two schemes by the spatial correlation measures for smaller gridcell distances, but worse by the temporal Spearman's rank correlation. Another observation is that the SS-CLM and SS-ANA are similar in spatial and temporal correlations. On the other hand, the SS-ANA and SS-ENS are better than the SS-CLM by the spatial similarity measure (Baddeley's delta) for large spatial scales. Furthermore, by the measure for capturing spatial variability, the SS-ENS is consistently better than the SS-ANA, which in turn is consistently better than the SS-CLM. On the whole, it is reasonable to conclude that the SS-ANA is a better choice over the SS-CLM. Between the SS-ANA and SS-ENS, we note that 1) the

SS-ANA depends on the choice of space–time domains for producing analogs, which may lead to boundary inconsistencies, whereas 2) the SS-ENS, on the other hand, is free of this issue, but is constrained by the ensemble size for optimal performance. In addition, for small-scale and light precipitation events, which are largely smoothed out in the statistics, it is observed that the SS-ENS yields more realistic-looking ensemble fields than the other two schemes. Ultimately, the choice of a proper SS scheme ought to be guided by the constraints and performance criteria of downstream applications such as hydrologic forecasting.

Acknowledgments. This work is supported by the Centralized Water Forecasting Demonstration program and the Advanced Hydrologic Prediction Service program of the Office of Water Prediction of the National Oceanic and Atmospheric Administration (NOAA), which we gratefully acknowledge. The first author would like to thank Tomislava Vukicevic of the NWS Office of Water Prediction for valuable suggestions. We thank the editor and the anonymous referees for their valuable comments that lead to improved content and clarity of this work.

APPENDIX

Rank Structure Analysis

Suppose a rank structure is being constructed for a number of space–time forecast points, each with an associated ensemble of size n. Given any two of these space–time forecast points, let P and Q denote their observed historical samples, each of size n, that provide the ranks to the rank structure. Let X be an ensemble (a sample from a predictive distribution) that receives the ranking from P. Let Y be the ensemble corresponding to Q. Furthermore, let R and S denote the rank variable for P and Q, respectively, and W and Z for X and Y, respectively.

If each of P, Q, X, and Y has distinct values, then once the SS-CLM is applied, X and Y will acquire the ranking of P and Q, respectively. As a result, the Spearman's rank correlation coefficient between X and Y will be equal to that between P and Q, since the Spearman's rank correlation coefficient between two variables is defined to be the Pearson correlation between their rank variables.

If there are identical values in the samples, a common practice in computing the Spearman's rank correlation is to assign fractional ranks equal to the average of their positions in ascending order of the values. To simplify the analysis, we assume that the number of zeroes in both P and Q is k. This assumption corresponds to the common experience that observed precipitation variables from adjacent locations have comparable PoPs. Next, we assume that the number of zeroes in both X and Y is l. If a large-scale precipitation event is predicted by an ensemble forecast, then it is often the case that there is a high fraction of positive members in the ensemble, which usually means that $l < k$. We ask how the Spearman's rank correlation $\gamma_{X,Y}$ between X and Y, will differ from $\gamma_{P,Q}$. Since $\gamma_{P,Q} = \rho_{R,S}$, the Pearson's correlation between R and S, and $\gamma_{X,Y} = \rho_{W,Z}$, we shall focus our attention on the computation of $\rho_{R,S}$ and $\rho_{W,Z}$. Let $\overline{(\cdot)}$ denote the mean value of a variable. We have

$$\rho_{R,S} = \frac{\sum_{i=1}^{n} r_i s_i - n\overline{R}\,\overline{S}}{\sqrt{\sum_{i=1}^{n} r_i^2 - n\overline{R}^2}\sqrt{\sum_{i=1}^{n} s_i^2 - n\overline{S}^2}}. \tag{A1}$$

Similarly, we have

$$\rho_{W,Z} = \frac{\sum_{i=1}^{n} w_i z_i - n\overline{W}\,\overline{Z}}{\sqrt{\sum_{i=1}^{n} w_i^2 - n\overline{W}^2}\sqrt{\sum_{i=1}^{n} z_i^2 - n\overline{Z}^2}}. \tag{A2}$$

Variable R has k identical fractional ranks equal to $(1 + \cdots + k)/k$ for the zeroes. A quick calculation yields

$$\overline{R} = \frac{1}{2}(1 + n),$$

which does not depend on k. We can readily see that $\overline{R} = \overline{S} = \overline{W} = \overline{Z}$. The terms $\sum_{i=1}^{n} r_i^2$, $\sum_{i=1}^{n} s_i^2$, $\sum_{i=1}^{n} z_i^2$, and $\sum_{i=1}^{n} w_i^2$ in the above equations, however, do depend on k or l, as it is easy to verify, for example, that

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} s_i^2 = \frac{1}{6}n(n+1)(2n+1) - \frac{1}{12}k(k+1)(k-1).$$

Since $k \neq l$, we have

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} s_i^2 \neq \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} w_i^2,$$

which means, in general, $\rho_{R,S} \neq \rho_{W,Z}$. Furthermore, for terms $\sum_{i=1}^{n} r_i s_i$ and $\sum_{i=1}^{n} w_i z_i$, randomness plays a role in their calculation. Indeed, if $l < k$, the smallest $k - l$ positive values in X and Y will be randomly placed in the positions corresponding to some zeroes in P and Q, respectively, by the SS-CLM. This can result in reduction of correlation in the processed ensembles. The above analysis shows that presence of tied ranks in the rank structures compromises the effectiveness of the SS schemes.

REFERENCES

Ajami, N. K., G. M. Hornberger, and D. L. Sunding, 2008: Sustainable water resource management under hydrological uncertainty. *Water Resour. Res.*, **44**, W11406, https://doi.org/10.1029/2007WR006736.

Baddeley, A., 1992: Errors in binary images and an L^p version of the Hausdorff metric. *Nieuw Arch. Wiskunde*, **10**, 157–183.

Berrocal, V. J., and A. Raftery, 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.

——, ——, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.*, **135**, 1386–1402, https://doi.org/10.1175/MWR3341.1.

Brown, J. D., 2015: An evaluation of the minimum requirements for meteorological reforecasts from the Global Ensemble Forecast System (GEFS) of the U.S. National Weather Service (NWS) in support of the calibration and validation of the NWS Hydrologic Ensemble Forecast Service (HEFS). Tech. Rep., 120 pp., http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/HSL_LYNT_DG133W-13-CQ-0042_SubK_2013_1003_Task_3_Deliverable_04_report_FINAL.pdf.

——, L. Wu, M. He, S. Regonda, H. Lee, and D.-J. Seo, 2014a: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *J. Hydrol.*, **519**, 2869–2889, https://doi.org/10.1016/j.jhydrol.2014.05.028.

——, M. He, S. Regonda, L. Wu, H. Lee, and D.-J. Seo, 2014b: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *J. Hydrol.*, **519**, 2847–2868, https://doi.org/10.1016/j.jhydrol.2014.05.030.

Chernick, M. R., 2008: *Bootstrap Methods: A Guide for Practitioners and Researchers*. 2nd ed., Wiley, 400 pp.

Clark, A. J., W. A. Gallus, 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.

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.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Daly, C., W. P. Gibson, M. Doggett, J. Smith, and G. Taylor, 2004: Up-to-date monthly climate maps for the conterminous United States. *14th Conf. on Applied Climatology*, Seattle, WA, Amer. Meteor. Soc., P5.1, https://ams.confex.com/ams/84Annual/techprogram/paper_71444.htm.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

——, 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, https://doi.org/10.1175/2009WAF2222251.1.

Eldardiry, H., E. Habib, Y. Zhang, and J. Graschel, 2015: Artifacts in Stage IV NWS real-time multisensor precipitation estimates and impacts on identification of maximum series. *J. Hydrol. Eng.*, **22**, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001291.

Gel, Y., A. E. Raftery, and T. Gneiting, 2004: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method. *J. Amer. Stat. Assoc.*, **99**, 575–583, https://doi.org/10.1198/016214504000000872.

Georgakakos, A. P., H. Yao, M. Mullusky, and K. P. Georgakakos, 1998: Impacts of climate variability on the operational forecast and management of the Upper Des Moines River basin. *Water Resour. Res.*, **34**, 799–821, https://doi.org/10.1029/97WR03135.

Gilleland, E., 2011: Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases. *Wea. Forecasting*, **26**, 409–415, https://doi.org/10.1175/WAF-D-10-05061.1.

Greene, D., and M. Hudlow, 1982: Hydrometeorologic grid mapping procedures. *Proc. Int. Symp. on Hydrometeorology*, Denver, CO, AWRA, 20 pp.

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

——, G. Bates, J. Whitaker, D. Murray, M. Fiorino, T. Galarneau, 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.

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.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jolliffe, I. T., and D. B. Stephenson, 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. Wiley, 292 pp.

Kelly, K. S., and R. Krzysztofowicz, 1997: A bivariate meta-Gaussian density for use in hydrology. *Stoch. Hydrol. Hydraul.*, **11**, 17–31, https://doi.org/10.1007/BF02428423.

Kitzmiller, D., and Coauthors, 2011: Evolving multisensor precipitation estimation methods: Their impacts on flow prediction using a distributed hydrologic model. *J. Hydrometeor.*, **12**, 1414–1431, https://doi.org/10.1175/JHM-D-10-05038.1.

Maglaras, G. J., J. S. Waldstreicher, P. J. Kocin, A. F. Gigi, and R. A. Marine, 1995: Winter weather forecasting throughout the eastern United States. Part 1: An overview. *Wea. Forecasting*, **10**, 5–20, https://doi.org/10.1175/1520-0434(1995)010<0005:WWFTTE>2.0.CO;2.

Marzban, C., and S. Sandgathe, 2010: Optical flow for verification. *Wea. Forecasting*, **25**, 1479–1494, https://doi.org/10.1175/2010WAF2222351.1.

Pinson, P., 2012: Adaptive calibration of (u,v)-wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1273–1284, https://doi.org/10.1002/qj.1873.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

R Core Team, 2016: R: A language and environment for statistical computing. R Foundation for Statistical Computing, https://www.R-project.org.

Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing 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.

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

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

Schuhen, N., T. T. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.*, **140**, 3204–3219, https://doi.org/10.1175/MWR-D-12-00028.1.

Schwedler, B. R. J., and M. E. Baldwin, 2011: Diagnosing the sensitivity of binary image measures to bias, location, and event frequency within a forecast verification framework. *Wea. Forecasting*, **26**, 1032–1044, https://doi.org/10.1175/WAF-D-11-00032.1.

Seo, D.-J., A. Seed, and G. Delrieu, 2011: Radar and multisensor rainfall estimation for hydrologic applications. *Rainfall: State of the Science*, *Geophys. Monogr.*, Vol. 191, Amer. Geophys. Union, 79–104.

Sklar, A., 1973: Random variables, joint distribution functions, and copulas. *Kybernetika*, **9** (6), 449–460.

Sonia, C.-G., L.-A. Emilio, and E.-M. M. Dolores, 2012: Selection of a plotting position for a normal Q-Q plot. R script. *J. Commun. Comput.*, **9**, 243–250.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.*, **110**, D08111, https://doi.org/10.1029/2004JD005395.

Voisin, N., J. C. Schaake, and D. P. Lettenmeier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 1603–1627, https://doi.org/10.1175/2010WAF2222367.1.

Vrac, M., and P. Friederichs, 2015: Multivariate—intervariable, spatial, and temporal—bias correction. *J. Climate*, **28**, 218–237, https://doi.org/10.1175/JCLI-D-14-00059.1.

Wilks, D. S., 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, S. Cong, J. D. Brown, 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.

Yang, X., S. Sharma, R. Siddique, S. J. Greybush, and A. Mejia, 2017: Postprocessing of GEFS precipitation ensemble reforecasts over the U.S. Mid-Atlantic region. *Mon. Wea. Rev.*, **145**, 1641–1658, https://doi.org/10.1175/MWR-D-16-0251.1.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, **105**, 10 129–10 146, https://doi.org/10.1029/1999JD901087.

Zhang, Y., S. Reed, and D. Kitzmiller, 2011: Effects of retrospective gauge-based readjustment of multisensor precipitation estimates on hydrologic simulations. *J. Hydrometeor.*, **12**, 429–443, https://doi.org/10.1175/2010JHM1200.1.