Accounting for Representativeness in the Verification of Ensemble Precipitation Forecasts

ZIED BEN BOUALLEGUE AND THOMAS HAIDEN

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

NICHOLAS J. WEBER

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

THOMAS M. HAMILL

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

DAVID S. RICHARDSON

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 30 September 2019, in final form 19 February 2020)

ABSTRACT

Spatial variability of precipitation is analyzed to characterize to what extent precipitation observed at a single location is representative of precipitation over a larger area. Characterization of precipitation representativeness is made in probabilistic terms using a parametric approach, namely, by fitting a censored shifted gamma distribution to observation measurements. Parameters are estimated and analyzed for independent precipitation datasets, among which one is based on high-density gauge measurements. The results of this analysis serve as a basis for accounting for representativeness error in an ensemble verification process. Uncertainty associated with the scale mismatch between forecast and observation is accounted for by applying a perturbed-ensemble approach before the computation of scores. Verification results reveal a large impact of representativeness error on precipitation forecast reliability and skill estimates. The parametric model and estimated coefficients presented in this study could be used directly for forecast postprocessing to partly compensate for the limitation of any modeling system in terms of precipitation subgrid-scale variability.

1. Introduction

The scale mismatch between in situ observations and gridded numerical weather prediction (NWP) forecasts is called representativeness error and is a challenge to be addressed in a number of applications (Göber et al. 2008; Janjić et al. 2018). For example, in forecast verification, skill estimates can differ substantially when the forecast is compared against its own analysis field or against point observations (Pinson and Hagedorn 2012; Feldmann et al. 2019). The presence of representativeness error in the latter case contributes to skill estimate differences. In more general terms, observation errors in forecast verification (the main topic of this paper) have gathered more attention as the accuracy of the forecast approaches the accuracy of observation measurements (Saetra et al. 2004; Candille and Talagrand 2008; Santos and Ghelli 2012; Röpnack et al. 2013; Massonnet et al. 2016; Jolliffe 2017; Ferro 2017; Duc and Saito 2018).

From the literature, we know that accounting for observation errors can have a large impact in the context of ensemble forecast verification, in particular when focusing on forecast reliability (Saetra et al. 2004; Candille and Talagrand 2008; Yamaguchi et al. 2016). Ensemble forecasts are a collection of forecasts valid for the same lead time that aim to capture the forecast uncertainty of the day (Leutbecher and Palmer 2008), and reliability is a desirable property for an ensemble forecast. Broadly speaking, a reliable ensemble forecast ensures statistical consistency between the dispersion of the ensemble (which represents the forecast uncertainty) and the forecast error

Corresponding author: Zied Ben Bouallegue, zied.benbouallegue@ecmwf.int

DOI: 10.1175/MWR-D-19-0323.1

with respect to the observations. If observation errors are not accounted for during the ensemble verification process, then the investigator may draw inappropriate conclusions about the quality of the prediction system. For example, suppose a coarser-resolution global ensemble appears (misleadingly) to be reliable with respect to point observations. With respect to verification against coarser gridded analyses, it may actually be overspread, indicating the potential for changes in the ensemble prediction system to provide less spread and potentially greater forecast resolution. Ultimately, dismissing observation errors in the verification process can have as an unfortunate consequence the inappropriate ranking of competing forecasting systems (Ferro 2017).

To account for observation uncertainty in the ensemble verification process, observation errors must first be characterized. This characterization is one objective of this paper with a focus on precipitation. Observation errors are the sum of measurement errors and representativeness errors.[1] In the following, we assume that representativeness error is the dominant contribution to observation errors associated with precipitation gauge measurements for our applications (Lopez et al. 2011). Representativeness of precipitation observations has already been investigated in previous studies, in particular in the framework of data assimilation (Lopez 2011, 2013). But here the focus is on daily precipitation (rather than short accumulation periods) point observations (rather than aggregated observations), and we apply a state-of the-art probabilistic parametric model.

The representativeness of precipitation observations can be described in probabilistic terms as the relationship between observations at two different spatial scales. Statistical models are used to estimate the properties of precipitation representativeness error and its peculiar characteristics: a probability distribution with a long tail and an uncertainty that grows with precipitation intensity. These statistical methods have been developed in the context of ensemble postprocessing to account for model limitation in representing subgrid variability and correct simultaneously for systematic forecast deficiencies such as biases (Wilks and Vannitsem 2018). Among others, successful methods encompass extended logistic regression (Wilks 2009; Ben Bouallègue 2013), quantile mapping (Hamill et al. 2017; Hamill and Scheuerer 2018), and nonhomogeneous regression (Scheuerer and Hamill 2015; Baran and Nemoda 2016). The latter approach, which relies on a parametric model based on the gamma distribution, is employed in this study. Because of its simplicity, the method followed here could be considered as a benchmark for more complex approaches that, for example, describe precipitation subgrid variability as a function of the weather situation (Pillosu and Hewson 2017).

The estimated uncertainty associated with precipitation measurements can be incorporated in the process that compares ensemble precipitation forecasts against synoptic station (SYNOP) observations.[2] Another objective of this paper is to assess the impact of accounting for observation representativeness on ensemble precipitation verification results. Practically, a perturbed-ensemble method is applied. It consists of adding observation uncertainty to the forecasts. A perturbation drawn from the parametric distributions is added to each ensemble member. The impact of this approach on verification results is illustrated when applied to the ensemble predictions generated at the European Centre for Medium-Range Weather Forecasts (ECMWF).

This paper is organized as follows: section 2 introduces the observation datasets, the parametric model used for the description of the relationship between observations at different scales, and the verification metrics for the validation of this approach. Section 3 presents details of the methodology and findings related to the characterization of the observation uncertainty. Section 4 describes how to account for observation uncertainty in the verification process and discusses the impact on verification results. Section 5 concludes and indicates another possible application of the methodology developed here.

2. Data, parametric model, and validation metrics

a. Data

Spatial representativeness of precipitation observations is analyzed based on three independent datasets, to check the consistency of estimates from differing data sources. We focus exclusively on 24-h accumulated precipitation. The amount of representativeness error will depend on the accumulation period, with less representativeness error for longer accumulation periods than for shorter ones. The three datasets correspond to

1) a high-density point observation dataset based on rain gauges (HDOBS),
2) radar observations from the Next Generation Radars (NEXRAD), and

[1] The reader is invited to refer to Janjić et al. (2018) for a discussion on representativeness definitions in the literature and to Tustison et al. (2001) for theoretical considerations on representativeness error.

[2] Note that the observation uncertainty estimates are independent of the forecast and so could be used in other applications.

3) short-range forecasts from the Model for Prediction Across Scales (MPAS).

The first two datasets are observations in the sense of instrument-derived measurements, the third dataset is an observation proxy, an NWP model output at short forecast range. Only the first dataset provides point measurements, and the two others are gridded datasets representing precipitation spatial averages.

The spatial coverage differs from one dataset to another as illustrated in Fig. 1. HDOBS covers parts of Europe, NEXRAD is limited to the conterminous United States with a spatial resolution of 4 km, while MPAS data are global with a grid spacing of 3 km. Similarly, the temporal coverage differs between datasets: different combinations of observing days are selected in each case as detailed below. This is acceptable for our study because the analysis of precipitation variability is performed for each dataset separately.

1) HDOBS

HDOBS corresponds to rain gauge measurements provided by ECMWF member and cooperating states in addition to the network of SYNOP observations (Haiden et al. 2018). The dataset covers Europe and the number of measurements varies from day to day. The station density relative to SYNOP is typically enhanced by a factor of 2–10. In this study, we use four nonconsecutive months of data (January, April, July, and October 2018) with on average around 5000 observations per day.

2) NEXRAD

NEXRAD is a U.S.-wide network of Weather Surveillance Radar-1988 Doppler (WSR-88D) instruments (Fulton et al. 1998). This study is based on a limited sample of observations days, corresponding to days 1, 8, 15, 22, and 29 of the months of June 2017, November 2017, and February 2018. Observations west of longitude 105°W are excluded from the analysis to reduce the impact of issues in the radar data that are due to orography.

3) MPAS

MPAS forecasts are based on a global nonhydrostatic model with a spherical Voronoi mesh (Skamarock et al. 2012). Precipitation is accumulated between forecast ranges 12 and 36 h on a uniform global mesh with 3-km grid spacing. The MPAS dataset comprises the three following days: 2 December 2003, 22 November 2011, and 8 February 2013. They correspond to case 1, case 2, and case 3 of a study on the predictive skill of subseasonal predictions in a global convection-permitting model (Weber and Mass 2019). Here, only midlatitudes (latitudes between 65° and 25°S and 25° and 65°N) are considered. Note that MPAS is not used for verification but as a dataset for estimating representativeness error characteristics. Characterization of forecast representativeness error as opposed to observation representativeness error is also beyond the scope of this study.

b. *Parametric model*

The parametric model of variability on unrepresented scales consists of fitting a censored, shifted gamma distribution (CSGD). It has been shown that the CSGD is well suited for describing precipitation probability distributions (Scheuerer and Hamill 2015). We recall here the formalism of this model before explaining how it is applied to the description of observation uncertainty.

The gamma distribution is a two-parameter distribution, with scale parameter k and shape parameter θ. The shift of the gamma distribution associated with a left censoring to 0 allows us to better represent the probability of no precipitation. The skewness of the gamma distribution depends only on its shape parameter θ. The two parameters k and θ are related to the mean μ and standard deviation σ of the gamma distribution by

$$k = \mu^2/\sigma^2 \quad \text{and} \quad \theta = \sigma^2/\mu. \tag{1}$$

The cumulative distribution function of CSGD (with left censoring at zero, denoted $\tilde{F}_{k,\theta,\delta}$) takes the form

$$\tilde{F}_{k,\theta,\delta}(y) = \begin{cases} F_k\left(\dfrac{y+\delta}{\theta}\right) & \text{for} \quad y \geq 0 \\ 0 & \text{for} \quad y < 0 \end{cases}, \tag{2}$$

where F_k is the cumulative distribution function of gamma distribution with unit scale and shape parameter k, and with $\delta > 0$, the shift parameters that controls the probability of zero precipitation (Scheuerer and Hamill 2015).

The CSGD is fitted in the form of a conditional distribution for observed precipitation at one spatial scale, say B, given the observed precipitation at a larger scale, say A. More precisely, we are interested in the conditional probability

$$P(Y_B|Y_A),$$

which is the probability of the random variable Y_B, representing the observation at smaller scale, given the random variable Y_A, representing the observation at a larger scale (e.g., the grid scale of an NWP model). We assume that this conditional distribution takes the parametric form described by a CSGD [Eq. (2)].

Exploratory analysis of the model sensitivity to the number of parameters suggests that five coefficients are

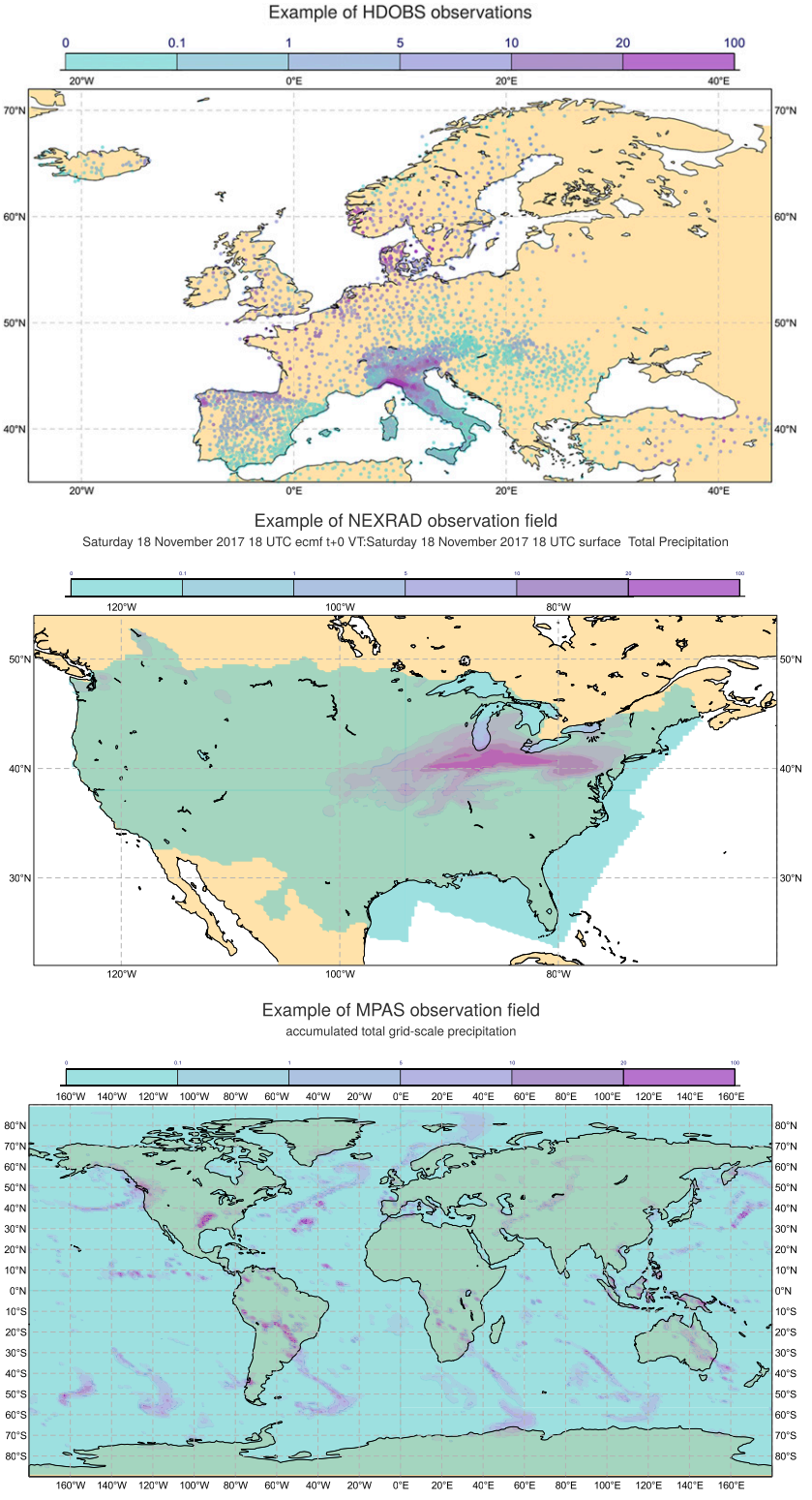

FIG. 1. Twenty-four-hour accumulated precipitation (mm) as seen by the three datasets used in this study: (top) HDOBS measurements on 1 Jan 2018, (middle) NEXRAD observation on 18 Nov 2017, and (bottom) MPAS short-range forecast on 22 Nov 2011.

required to describe this distribution of Y_B accurately for the three studied datasets. Two coefficients (α_0 and α_1) are associated with the mean of the distribution μ_B:

$$\mu_B(y_A) = \alpha_0 + \alpha_1 y_A, \tag{3}$$

which is a function of the observed precipitation at scale A (y_A). Two other coefficients (β_0 and β_1) are associated with the standard deviation of the distribution σ_B:

$$\sigma_B(y_A) = \beta_0 + \beta_1 (y_A)^{1/2}, \tag{4}$$

which is a function of the square root of the observed precipitation at scale A [$(y_A)^{1/2}$]. The use of a power transformation in the relationship between precipitation intensity and uncertainty can be traced back to pioneering work on postprocessing of ensemble precipitation forecasts (Hamill et al. 2008). The fifth coefficient corresponds to δ, which defines the shift associated with the CSGD.

The five distribution parameters (α_0, α_1, β_0, β_1, and δ) are estimated by minimizing the mean continuous ranked probability score (CRPS) over a test sample (see section 3a). Following Gneiting and Raftery (2007), the CRPS is defined for a distribution $F(y_A)$ and an observation y_B as follows:

$$\text{CRPS} = E_X|X - y_B| - 0.5E_{X,X'}|X - X'|, \tag{5}$$

where X and X' are independent random variables with distribution $F(y_A)$. In the case in which F takes the form of a CSGD [Eq. (2)], the CRPS can be expressed in a closed form [see Eq. (10) in Scheuerer and Hamill 2015]. Optimization is performed using squared parameters to ensure that they are positive, and with $\alpha_0 = 0.1$, $\alpha_1 = 1$, $\beta_0 = 0.1$, $\beta_1 = 1$, and $\delta = 0.1$ as initial values of the optimization process.

c. PIT histograms

The validity of the parametric method described in the previous section is checked by means of probability integral transform (PIT; Raftery et al. 2005) histograms. We apply the following diagnostic procedure: we consider percentiles associated with the CSGD for each element of the test sample. Percentiles are derived for equidistant probability levels ranging from 5% to 95% with a 5% interval. The rank of the observations when pooled with the distribution percentiles is aggregated and reported on a histogram.

PIT histograms are interpreted in the same way as rank histograms (Hamill and Colucci 1997), where a histogram close to a uniform distribution indicates reliability. PIT (or rank) histograms can be summarized in a single number (Wilks 2019). In the following, we compute the reliability index (RI; Monache et al. 2006):

$$\text{RI} = \frac{1}{m}\sum_{i}^{m+1}\left|\zeta_i - \frac{1}{m+1}\right|, \tag{6}$$

where $m + 1$ is the number of equally sized bins and ζ_i is the frequency of observations in the ith bin. RI takes a minimum value of 0 when the system is perfectly calibrated. In addition, we assess reliability with an entropy measure (ψ; Taillardat et al. 2016):

$$\psi = \frac{-1}{\log(m+1)}\sum_{i=1}^{m+1}\zeta_i \log(\zeta_i), \tag{7}$$

which takes an optimum value of 1 when the system is perfectly reliable and the sample size is infinite.

3. Observation representativeness error

In this section, observation errors are investigated based on the datasets and parametric model presented in section 2. An observation error is meant here as an outcome of the representativeness uncertainty associated with smaller-scale measurements with respect to observations averaged at a larger scale. Fundamentally, the aim is to characterize the relationship between precipitation averaged over an area A and precipitation measurements at B, where B is a point within the area A. This characterization defines the representativeness error associated with point observations (such as SYNOP measurements) and is used later in a forecast verification process (see section 4).

a. Method

With HDOBS data, point measurements can be directly compared with areal observations. For this, we consider single observations (denoted y_B) and regularly spaced neighborhood areas A defined as square areas with length Δ_A. When at least five observations are found within an area, the averaged precipitation is computed and is denoted y_A. Repeating this for all days of the dataset, we obtain a sample of pairs (y_A, y_B). This is done separately for each month of the dataset. We note that there is an uncertainty associated with this method because we do not know the actual value of y_A but rather just an estimate that is based on a limited set of point observations.[3]

An example of a sample (y_A, y_B) from the HDOBS dataset is provided in Fig. 2 considering a neighborhood size $\Delta_A = 20$ km. It shows how point observations (y_B)

[3] Increasing the minimum number of observations per grid box has little impact on the results.

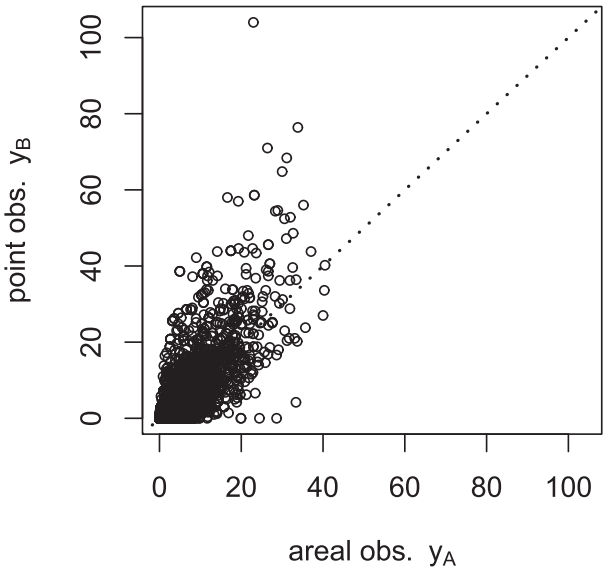

FIG. 2. Example of a sample of pairs (y_A, y_B) where y_A are averaged observations over areas of 20 km by 20 km and y_B are point observations within the corresponding areas; y_A and y_B units are millimeters per 24 hours. Observations are from the HDOBS dataset for July 2018.

do not always coincide with averaged precipitation within a surrounding area (y_A). Zero precipitation can be observed at one specific location while the areal precipitation can be large. There are also cases where point measurements are large and areal observations are much smaller. This illustrates the mismatch between areal and point observations that we aim to characterize.

The next step is to fit the parametric model with the pairs (y_A, y_B) to describe in probabilistic terms the relationship between these two quantities. Parameters of the CSGD are estimated for a range of neighborhood sizes Δ_A: 10, 20, . . . , 140, 150 km. The parameters are estimated for each month separately, considering randomly selected blocks of days within each month. The parameter estimates are aggregated over the four months by computing the median. By repeating the process 1000 times, with a new random selection of days each time, one obtains a distribution of parameter estimates from which one can derive confidence intervals. Eventually, the estimated parameters ($\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\delta}$), as well as the corresponding uncertainty can be plotted as a function of the size of the averaging area Δ_A.

For the two other datasets (NEXRAD and MPAS), we infer the uncertainty associated with point measurements based on gridded observations using an extrapolation technique. For each of these datasets, we proceed as follows:

- we randomly select a square area A with sides of length Δ_A within the data domain,
- we average the grid values within this area and denote it y_A,
- we randomly select a square area B with sides of length Δ_B within the area A, and
- we average the available measurements within this area B and denote it y_B.

We repeat these steps 1000 times to obtain a large sample of pairs (y_A, y_B).

This time, the model parameters are estimated for a range of Δ_A and Δ_B with $\Delta_B < \Delta_A$. For NEXRAD, with a native grid spacing of 4 km, we use Δ_A and Δ_B in {4, 12, 20, 28, 36, 44, 52, 60, 68, 76, 84, 92, 100, 108, 116, 124, 132, 140, 148} km. For MPAS, with a native grid spacing of 3 km, we use Δ_A and Δ_B in {3, 12, 24, 36, 48, 60, 72, 84, 96, 108, 120, 132, 144} km. To extrapolate the parameters for $\Delta_B = 0$, we fix the scale of A (Δ_A) and consider each estimated parameter as a function of the scale of B (Δ_B). Using a second-order approximation, we can represent this function as a quadratic polynomial in Δ_B, and finally infer the parameter value for any Δ_B.

The extrapolation procedure is illustrated in Fig. 3 for β_1, the most critical parameter of the CSGD model. The parameters are estimated for all possible combinations of Δ_A and Δ_B defined above, with the restriction $\Delta_B < \Delta_A$. The results are presented as a function of Δ_B, where each line corresponds to a different larger-scale A. As we can see, this function, $\hat{\beta}_1(\Delta_B)$, can be approximated by a second-order polynomial for each Δ_A. For a given size Δ_A, the value of the function for $\Delta_B = 0$ is represented by a point: it corresponds to the extrapolated value of the parameter $\hat{\beta}_1$ associated with point measurements. Eventually, we obtain, for each model parameter, the extrapolated values at $\Delta_B = 0$ for a range of averaging areas Δ_A.

In our analysis, the target scale corresponds to point observations. However, one might be interested, for example, in representativeness uncertainty of a gridded precipitation dataset. In that case, the method followed here could easily be adapted by setting the target scale Δ_B appropriately.

b. *Parametric model results*

Figure 4 is central to this study. The estimated CSGD parameters ($\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\delta}$) are plotted as a function of the size of the averaging area Δ_A. The results are based on the HDOBS dataset. Extrapolated model parameters based on NEXRAD and MPAS are also plotted, showing an overall good agreement between the coefficients derived from the three independent datasets.

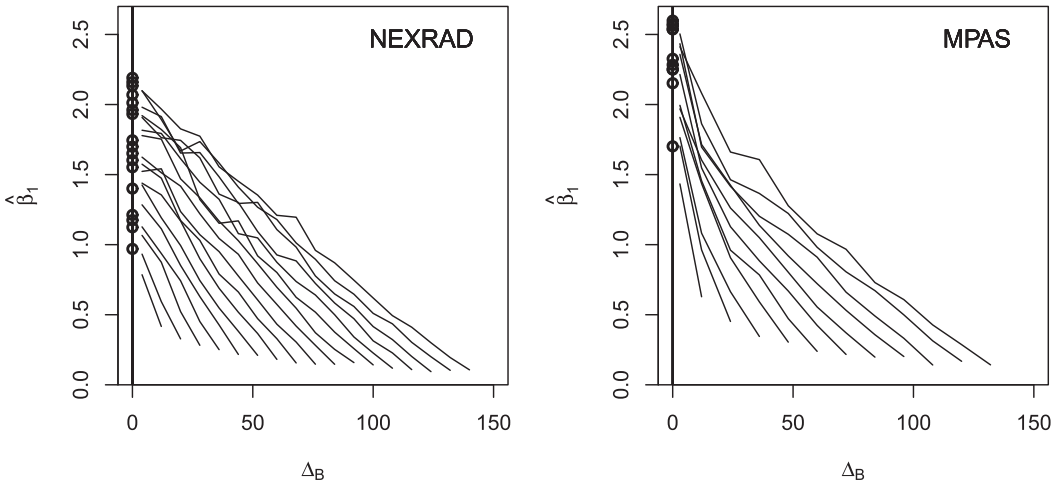

FIG. 3. For (left) NEXRAD and (right) MPAS, estimated parameters $\hat{\beta}_1$ for different combinations of smaller scale (Δ_B; x axis) and larger scales (Δ_A). From left to right, successive lines correspond to increasing value of Δ_A (see the text for the exact values). Values for point observations ($\Delta_B = 0$) are extrapolated and are represented by circles.

This figure can be used as follows: for a given model, representativeness uncertainty associated with the corresponding grid spacing Δ_A can be directly inferred from the red curve. So, Fig. 4 provides all required parameter values for accounting for precipitation observation representativeness error using our CSGD model.

A short description along with an interpretation of the results is now provided. On the one hand, the estimated coefficients associated with the mean of the distribution ($\hat{\alpha}_0$ and $\hat{\alpha}_1$) do not vary substantially with the averaging scale Δ_A. The multiplicative parameter $\hat{\alpha}_1$ is constant around 1, while the additive parameter $\hat{\alpha}_0$ is around 0.1 mm $(24\,\text{h})^{-1}$ for averaging scales greater than 20 km. Similarly, the shift parameter $\hat{\delta}$ is small and exhibits values that are comparable to those of $\hat{\alpha}_0$. So, the mean of the CSGD is generally close to y_A, which means that the expected mean precipitation intensity does not vary across scales.

On the other hand, one of the two coefficients associated with the variance of the distribution ($\hat{\beta}_1$) exhibits larger variability across scales than the other coefficients. Parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ influence the uncertainty associated with the CSGD distribution. Indeed, they determine the variance and skewness of the distribution through the shape parameter θ [Eq. (1)]. In particular, $\hat{\beta}_1$ brings heteroscedasticity into the model: it allows the precipitation uncertainty to be a function of the precipitation intensity. As expected, representativeness error of a single observation increases with the size of the target grid box, in agreement with results from previous studies (Lopez et al. 2011).

The overall good agreement between the coefficients estimated from HDOBS and the two other datasets is noticed for four of the five parameters. Parameter $\hat{\alpha}_0$ is smaller for NEXRAD, but MPAS has similar values to HDOBS. Parameter $\hat{\alpha}_1$ is smaller than HDOBS for both NEXRAD and MPAS and fluctuates between 0.9 and 1. Parameter $\hat{\beta}_0$ appears to be similar for all three datasets. δ is small, with the estimates for HDOBS being larger

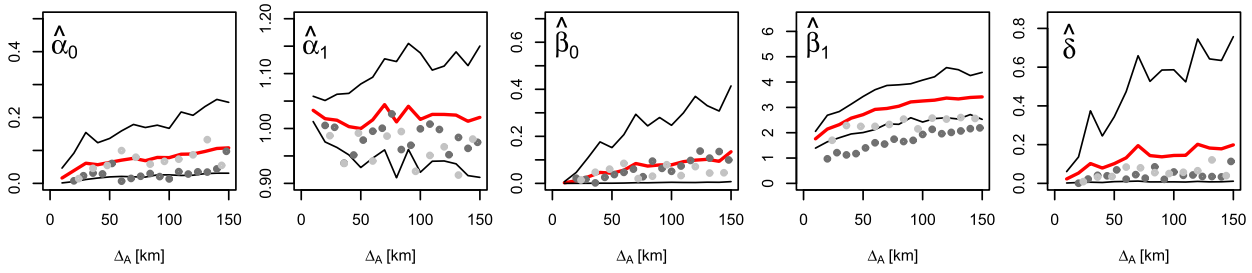

FIG. 4. Estimated parameters for the CSGD model, $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\delta}$, as a function of the averaging scale Δ_A (km) using HDOBS (red lines). Black lines indicate the 5%–95% confidence intervals estimated with block bootstrapping. Extrapolated parameters for $\Delta_B = 0$ from the NEXRAD (dark gray points) and MPAS (light gray points) datasets are also reported (see the text).

than for both the other datasets. For these four coefficients, estimates based on NEXRAD and MPAS are generally within the uncertainty margins associated with the estimates based on HDOBS. The main difference is observed for one coefficient influencing the distribution variance: $\hat{\beta}_1$. This parameter is clearly smaller for NEXRAD and MPAS and converges for the two datasets at larger scales Δ_A. For smaller scales, MPAS-based estimates are closer to HDOBS-based estimates and in any case within their uncertainty margins while NEXRAD-based estimates are below the 5% confidence intervals. NEXRAD and MPAS are gridbox-averaged precipitation products, but areal precipitation is derived from a limited number of point observations (minimum 5) in the HDOBS case. This could explain partially the larger multiplicative coefficient $\hat{\beta}_1$ for HDOBS with respect to the two other datasets.

The estimated parameters characterize observation representativeness error throughout the year. However, seasonality in the magnitude of representativeness error should be expected (Lopez et al. 2011). Recently, a nonparametric approach for precipitation postprocessing has been proposed that incorporates meteorological conditions as an additional source of information for a better representation of precipitation subgrid-scale variability (Pillosu and Hewson 2017). Following this idea, future refinement of the present method could for example consider CSGD parameters that vary as a function of season, orography, or region (e.g., tropics vs extratropics).

c. Model validation

The parametric model is validated by means of PIT histograms and derived statistics (section 2c). We check whether the fitted model, a function of the observation at scale A (y_A), appropriately captures the observation at scale B (y_B). For this purpose, each sample of pairs (y_A, y_B) is split in two subsamples of equal size. Parameters are estimated on the first half while the second half is used to assess the reliability of the model. Random selection is applied for the partitioning of the dataset.

For each smaller-scale observation y_B, 19 equally spaced quantiles (with probability level 5%, 10%, ..., 90%, 95%) are derived from the fitted CSGD driven by the larger-scale observation y_A. The rank of y_B within the set of quantiles for each pair (y_A, y_B) is aggregated and displayed in the form of a histogram. Figure 5 shows the resulting PIT histogram for HDOBS for an area with length $\Delta_A = 30$ km. The derived statistics, reliability index (RI) and entropy ψ, are indicated on the histogram. PIT histograms for the other datasets and averaging scales Δ_A are summarized by these two summary statistics (not shown). In Fig. 5 (left), the histogram looks mostly flat, which indicates good reliability of the model. Similar values of RI and ψ are obtained for other datasets and aggregation scales, RI being in any case below 0.2 and ψ always exceeding 0.99 (not shown).

A further check consists of a visual inspection of quantile–quantile (Q–Q) plots for random draws of the parametric distribution and a set of points observations. The Q–Q plots help in diagnosing whether the two sets of precipitation amounts are drawn from the same marginal distribution. An example is provided in Fig. 5b. Here again, good agreement is visible between model output and observations. Relative to the Q–Q plot for the original sample pairs (gray points), the fitted model better captures the marginal distribution of the point measurements, in particular its tail. In other words, the CSGD approach allows us to generate large precipitation amounts at a more appropriate frequency.

4. Ensemble forecast verification

In this section, we assess the performance of 24-h precipitation forecasts from ECMWF's medium-range ensemble forecasts (ENS). For illustration purposes, we choose a verification period covering summer (June, July, and August) 2018 and a verification domain corresponding to the European area shown in the top panel of Fig. 1. Rain gauge measurements from SYNOP observations and the corresponding nearest grid-point forecasts are used in the verification process. The horizontal grid spacing of ENS forecasts is about 18 km. From Fig. 4, we read the estimated CSGD parameters for this averaging scale ($\Delta_A = 18$ km):

$$\hat{\alpha}_0 = 0.02, \quad \hat{\alpha}_1 = 1.0, \quad \hat{\beta}_0 = 0.0, \quad \hat{\beta}_1 = 2.0, \quad \text{and} \quad \hat{\delta} = 0.02. \tag{8}$$

These parameters allow us to make the link between point measurements and averaged precipitation at the model scale. This information is now used to account for observation uncertainty in the verification process.

a. Perturbed-ensemble approach

To account for observation error in the verification process of ensemble forecasts, we apply the so-called perturbed-ensemble approach that consists of convolving the forecast and observation error distributions (Anderson 1996; Saetra et al. 2004; Candille and Talagrand 2008). This approach leads to scoring rules that favor forecasts of the truth, and it is therefore recommended as a generic method to be applied in the presence of observation errors (Ferro 2017).

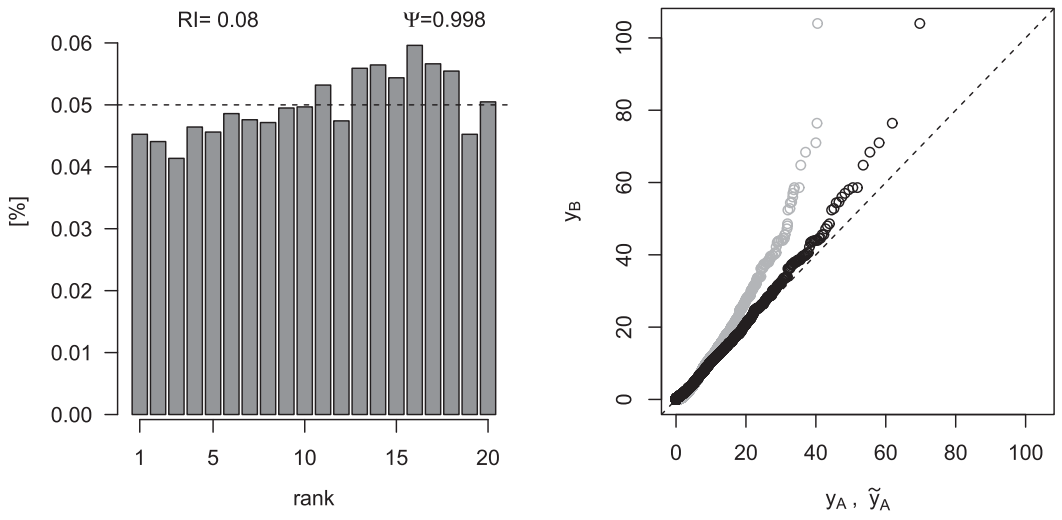

FIG. 5. (a) Model validation through PIT histograms and derived statistics. The dashed line indicates perfect calibration; (b) Q–Q plots [mm $(24\,\mathrm{h})^{-1}$] for the original sample (areal observations y_A), displayed in gray, and the fitted model (derived point observations $\tilde{y}_A$), displayed in black (right), with respect to the original point observations y_B. Results are for averaged observations for $\Delta_A = 30$ km and point observations ($\Delta_B = 0$ km) using the HDOBS dataset.

In practical terms, random noise is somehow added to the forecasts. Each ensemble member gets assigned a random value drawn from the fitted parametric distribution whose scale and shape parameters are a function of the original forecast value: the distribution is centered over the forecast value and its spread accounts for representativeness uncertainty. Combining Eqs. (3), (4), and (8), the scale and shape parameters of the CSGD, k and θ in Eq. (1), take the form

$$k = \frac{(0.02 + x)^2}{4x} \quad \text{and} \quad \theta = \frac{4x}{0.02 + x}, \tag{9}$$

where x is the forecast value of a single ensemble member.

This approach can also be seen as a forecast downscaling that provides a description of the subgrid-scale uncertainty that is not captured by the NWP model. The additional uncertainty from the perturbed-ensemble approach is merged with the original forecast uncertainty generated by the ensemble system, and together they represent the forecast uncertainty at the observation scale. In terms of ensemble spread, as measured by the standard deviation of the ensemble members with respect to the ensemble mean, observation uncertainty represents on average up to 45% of the total spread for a forecast range of 1 day, and this ratio decreases to reach a plateau around 15% for longer forecast lead times (after day 10).

An illustration of the observation uncertainty associated with single forecasts is provided in Fig. 6. For example, one member originally takes a value of 10 mm $(24\,\mathrm{h})^{-1}$ as indicated by a black vertical dashed line. Accounting for observation uncertainty, this member is assigned a value drawn from the distribution depicted in black when entering the verification process. Comparing this example with the case of a forecast with original value 0.5 (light gray) or 2 mm $(24\,\mathrm{h})^{-1}$ (dark gray), we see that the parametric model allows us to adjust the level of uncertainty as a function of the precipitation intensity.

b. *Verification metrics*

The impact of accounting for observation uncertainty in the verification process is assessed by focusing mainly on binary events. Two complementary verification tools are used: the reliability diagram, and the relative operating characteristic (ROC) curve (Wilks 2006). The former focuses on forecast reliability, that is the ability of the ensemble to capture the observation variability, while the second focuses on the discrimination ability of the forecast, that is its ability to distinguish between event and nonevent. Numerically, ROC curves are generated using the 51 possible probability thresholds issued by the 50-member ensemble.

In terms of summary performance measures, the Brier score (BS; Brier 1950) and the diagonal elementary score (DES; Ben Bouallègue et al. 2018, 2019) are applied in the form of skill scores. The verification sample climatology is used for the computation of DES as well as for the computation of skill scores considering the climatological probability of occurrence as a reference forecast. We also compute a general measure of

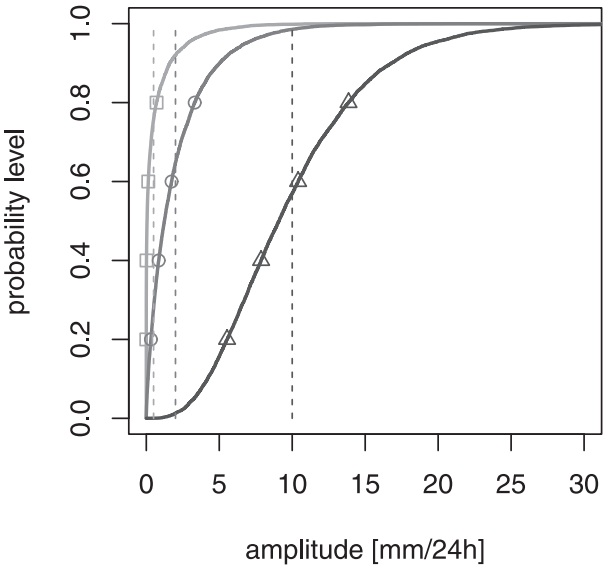

FIG. 6. Illustrative example of the perturbed-ensemble approach. Observation uncertainty is accounted for by replacing the forecast by a draw from the associated parametric distribution (represented here by its cumulative distribution function). Examples for forecasts of value 0.5 (light gray; squares), 2 (dark gray; circles), and 10 (black; triangles) mm $(24\,\mathrm{h})^{-1}$ as indicated by the dashed vertical lines and corresponding uncertainty distributions represented by the full lines.

ensemble performance for continuous variables, namely, the CRPS [see Eq. (5)]. For all verification metrics, block bootstrapping with blocks of 3 days and 1000 iterations is used to estimate confidence intervals. Applied to pairwise differences, confidence intervals including 0 indicate nonsignificant differences when accounting for representativeness.

c. *Impact on verification results*

The general impact of accounting for observation uncertainty is shown in Fig. 7: CRPS (the smaller the better), and relative CRPS difference, are plotted as a function of the forecast lead time. Results with and without observation uncertainty are compared. A large impact is visible in particular at short lead times: from 12% at day 1, the relative difference becomes no more statistically significant after day 7 (Fig. 7). Since the ensemble spread (and forecast error) is limited at the beginning of the forecast range, the scale mismatch between model and observations plays a substantial role. This is less the case at longer ranges when the ensemble spread (and forecast error) is larger.

Similar plots are provided in Fig. 8 but focusing this time on a 1 mm $(24\,\mathrm{h})^{-1}$ threshold event. BS and relative BS difference indicate a larger impact in this case (small event threshold) with up to 22% relative difference at short lead time (day 1). The impact on the Brier score is indeed particularly important for small thresholds as discussed below. Significant BS differences persists at longer lead times. Moreover, in absolute terms, BS exhibits a continuous increase (i.e., decrease in skill) with lead time when accounting for observation uncertainty (black line, Fig. 8a). This is not the case for the original results (gray line) but it is consistent with the expected error growth of the forecast. A similar behavior, although a bit weaker, can be seen in the CRPS.

The differences in terms of skill as measured by CRPS and BS can be explained by the large improvement of the forecast reliability when accounting for observation uncertainty. The ensemble spread is essentially increased by the perturbed-ensemble approach and, as a consequence, the perturbed-ensemble forecast is able to better capture the variability of point observations. To illustrate this point, reliability curves are shown considering two event thresholds: 1 mm $(24\,\mathrm{h})^{-1}$ in Fig. 9a and 20 mm $(24\,\mathrm{h})^{-1}$ in Fig. 10a. With observation uncertainty, the reliability curve is closer to the diagonal in the former case, but the impact appears to be mostly neutral in the latter case. The noticeable impact of the perturbed-ensemble approach on the ensemble reliability, as assessed here by reliability curves, is consistent with results from previous studies (Saetra et al. 2004; Candille and Talagrand 2008). The lack of reliability of high threshold forecast probabilities even after addressing representativeness provides evidence that there are remaining system problems that likely need to be addressed through prediction system improvement and/or postprocessing.

Now, we inspect the role of observation uncertainty when assessing ensemble forecasts in terms of discrimination. Figures 9c and 10c show the impact of accounting for observation errors on ROC curves. For an event with a 1 mm $(24\,\mathrm{h})^{-1}$ threshold, the impact is neutral: the two curves that are compared are on top of each other (Fig. 9c). The information content of the forecast is not modified for this type of event when adding the observation uncertainty to the forecast. However, when focusing on larger event thresholds, such as 20 mm $(24\,\mathrm{h})^{-1}$, the area under the two curves clearly differs in terms of extent (Fig. 10c). Note that the same number of members and so the same number of probability thresholds are considered in both cases. So, the perturbed-ensemble approach seems to produce a "shift" in probability distribution that appears beneficial for users with small probability thresholds. The ability of the perturbed ensemble to forecast large precipitation amounts, and so to capture the tail of the observation distribution, is rewarded in terms of forecast discrimination. The increase of ROC area estimates with the perturbed-ensemble approach also confirms results from a previous

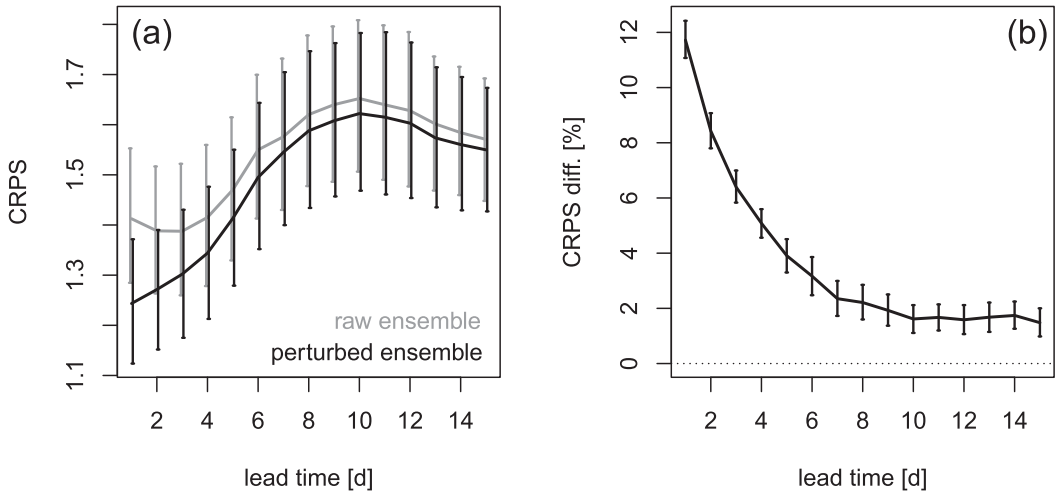

FIG. 7. (a) CRPS [mm $(24\,\mathrm{h})^{-1}$] as a function of the forecast lead time with (black) and without (gray) accounting for observation uncertainty, and (b) the corresponding CRPS relative difference (%). Vertical bars indicate 5%–95% confidence intervals.

study that focused on 850-hPa temperature forecasts (Candille and Talagrand 2008).

The sharpness diagram, which is usually included in reliability diagrams, is here plotted separately. In Figs. 9b and 10b, sharpness diagrams present the frequency of occurrence associated with each forecast probability level. Sharpness is not a measure of forecast skill per se, but this forecast attribute helps diagnose the impact of the perturbed-ensemble approach on the probabilistic forecast. In general, increasing the ensemble spread reduces forecast sharpness. For example, for low threshold events, when including observation uncertainty, we see a decrease of the frequency of high-probability forecasts, generally associated with overconfidence in a traditional verification framework, and hence an improvement of the forecast reliability. For high-threshold events, the number of forecasts with small probability values (falling in the 10%–30% bins) is increased, which allows us to capture more events (see ROC curve results), but this increase in small probability frequency does not improve the reliability of the forecast.

Figure 11 provides a summary of the forecast performance at day 5 as a function of the event threshold. In terms of Brier skill score (BSS), large impact is noted for low-intensity events, while in terms of the diagonal elementary skill score (DESS), larger differences are visible for more high-intensity events. This result is consistent with plots in Figs. 9 and 10, and with general

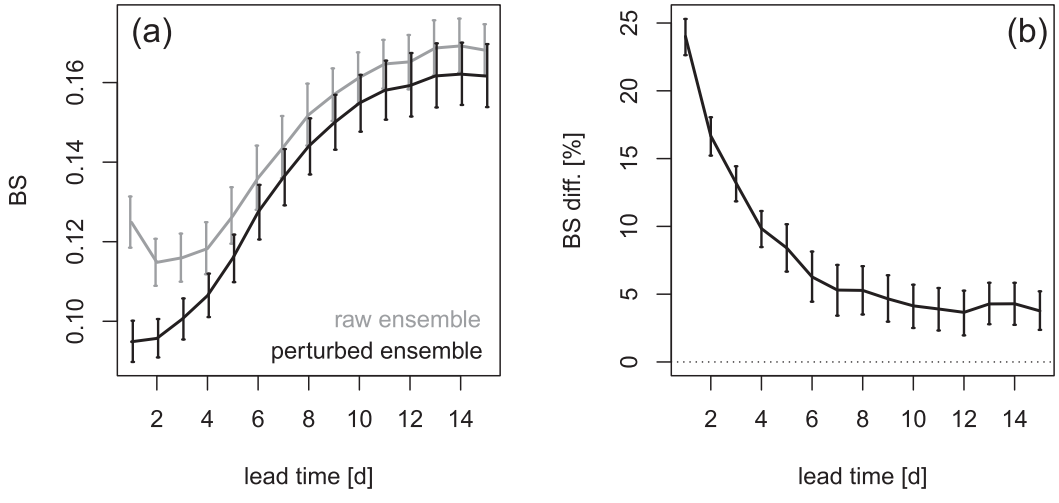

FIG. 8. As in Fig. 7, but for BS considering an event threshold of 1 mm $(24\,\mathrm{h})^{-1}$.

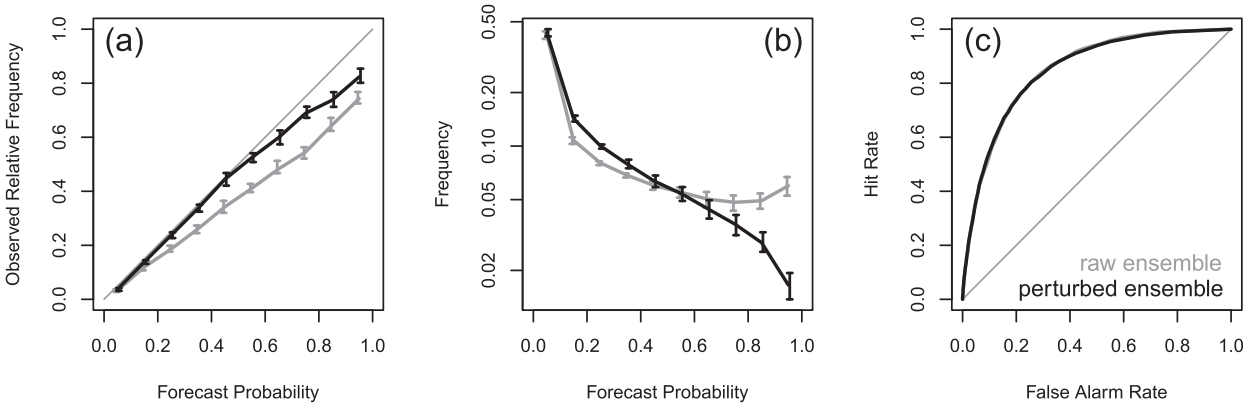

FIG. 9. (a) Reliability diagram, (b) sharpness diagram, and (c) ROC curve for an event threshold of 1 mm $(24\,\text{h})^{-1}$. Results are shown with (black) and without (gray) accounting for observation uncertainty when verifying ensemble precipitation forecasts at day 5. Vertical bars indicate 5%–95% confidence intervals.

characteristics of BS and DES: BS is more sensitive to reliability while DES is more sensitive to discrimination (Ben Bouallègue et al. 2019).

The impact on DESS for high-intensity events suggests that accounting for observation uncertainty could be crucial when assessing forecast skill for high-impact events. When the focus is exclusively on extreme events, that is, on the tail rather than on the whole distribution, an accurate estimation of the skill in the presence of observation uncertainty would probably benefit from a more pertinent model definition with the use, for example, of parametric distributions based on extreme value theory (Friederichs 2010).

To assess the robustness of our results, the sensitivity of the verification results to the specification of observation uncertainty is also investigated. For this purpose, we consider observation uncertainty specification based only on the analysis of the NEXRAD and MPAS datasets. The CSGD parameters estimated using these two datasets differ mainly in terms of $\hat{\beta}_1$ with respect to the original set of parameters based on the analysis of the HDOBS dataset (see Fig. 4). So, we can simply set $\hat{\beta}_1 = 1.5$ in Eq. (8) and recompute scores. With this parameter setting, the CRPS relative difference before and after accounting for representativeness drops to 9% (instead of 12%) at day 1 and to around 1% (instead of 2%) at day 10 (not shown). There are also quantitative differences in terms of BS and DES, but the qualitative impact is the same when using a more conservative set of parameters. As a further step, one could include in the verification results uncertainty associated with the estimation of the observation uncertainty itself.

5. Conclusions

This paper provides a general method for accounting for observation uncertainty when verifying ensemble precipitation forecasts. First, a parametric model based on a censored shifted gamma distribution is fitted to describe the representativeness error associated with point precipitation observations. This model, which provides a link between representativeness error and precipitation intensity, is

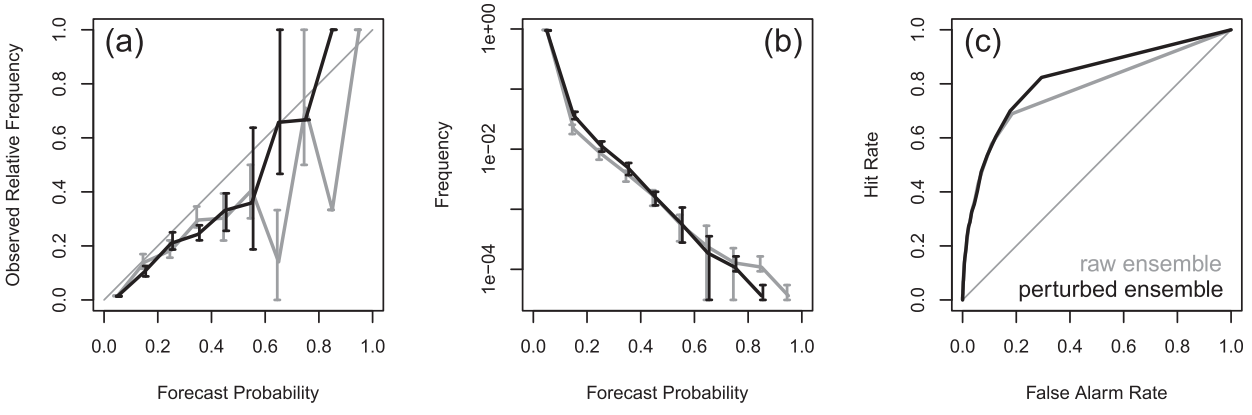

FIG. 10. As in Fig. 9, but for an event threshold of 20 mm $(24\,\text{h})^{-1}$.

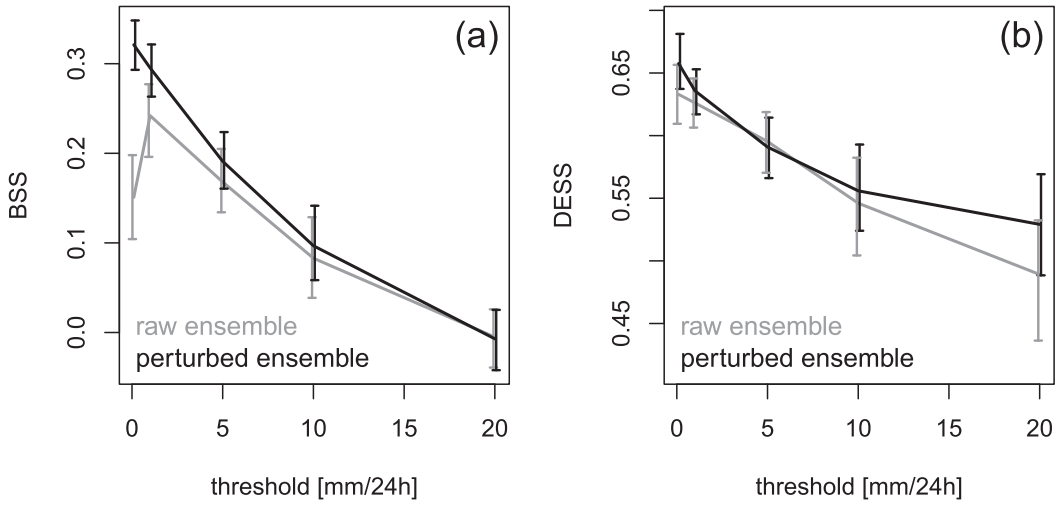

FIG. 11. (a) BSS and (b) DESS as a function of the event thresholds. Results are shown with (black) and without (gray) accounting for observation uncertainty when verifying forecast at day 5. Vertical bars indicate 5%–95% confidence intervals.

successfully validated by means of PIT histograms and Q–Q plots. Second, a perturbed-ensemble approach is applied: it consists of perturbing each ensemble member by means of this parametric model. This step allows us to include the uncertainty associated with station measurements in the verification process. Verification results derived with and without the perturbed-ensemble approach are compared and analyzed. It is shown, in summary, that accounting for observation representativeness error can have a large impact on the assessment of forecast reliability, forecast skill at short lead times, and potentially on forecast discrimination ability for high-intensity events.

An important side benefit of this study is that it provides the basis for a model-independent postprocessing method. Ensemble members (or deterministic forecasts) can be dressed with a CSGD using the parameters estimated in this study. The model fitting is based on independent observations only and so can be applied to forecasts from any model, simply adapting the parameters as a function of the model grid spacing. The derived probabilistic forecasts could be interpreted as valid at any given location of a model grid box. The proposed approach is fully parametric and, as such, is straightforward to apply (by any direct model-output user or forecast provider) to generate as many ''members'' as desired. This postprocessing can be seen as a way to account for model limitations that are due to subgrid-scale uncertainty, but it cannot correct for model-specific deficiencies.

The method proposed here is simple in its formulation and relies on only five parameters to describe observation uncertainty in general situations. This model could be developed further by considering that subgrid-scale uncertainty is weather-situation-dependent or by accounting for subgrid-scale orographic effects. For example, coefficients could vary with season and/or observation location. A more complex model would benefit both verification and potential postprocessing applications.

Acknowledgments. Fruitful discussions with Philippe Lopez and Martin Leutbecher are gratefully acknowledged. Pertinent and constructive comments from three anonymous reviewers allowed us to improve the paper.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, https://doi.org/10.1002/env.2391.

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.

——, T. Haiden, and D. S. Richardson, 2018: The diagonal score: Definition, properties, and interpretations. *Quart. J. Roy. Meteor. Soc.*, **144**, 1463–1473, https://doi.org/10.1002/qj.3293.

——, L. Magnusson, T. Haiden, and D. S. Richardson, 2019: Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quart. J. Roy. Meteor. Soc.*, **145**, 1741–1755, https://doi.org/10.1002/qj.3523.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 959–971, https://doi.org/10.1002/qj.268.

Duc, L., and K. Saito, 2018: Verification in the presence of observation errors: Bayesian point of view. *Quart. J. Roy. Meteor. Soc.*, **144**, 1063–1090, https://doi.org/10.1002/qj.3275.

Feldmann, K., D. Richardson, and T. Gneiting, 2019: Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophys. Res. Lett.*, **46**, 7744–7751, https://doi.org/10.1029/2019GL083189.

Ferro, C., 2017: Measuring forecast performance in the presence of observation error. *Quart. J. Roy. Meteor. Soc.*, **143**, 2665–2676, https://doi.org/10.1002/qj.3115.

Friederichs, P., 2010: Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, **13**, 109–132, https://doi.org/10.1007/s10687-010-0107-5.

Fulton, R., J. Breidenbach, D.-J. Seo, D. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting*, **13**, 377–395, https://doi.org/10.1175/1520-0434(1998)013<0377:TWRA>2.0.CO;2.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Göber, M., E. Zsoter, and D. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteor. Appl.*, **15**, 359–365, https://doi.org/10.1002/met.78.

Haiden, T., and Coauthors, 2018: Use of in situ surface observations at ECMWF. ECMWF Tech. Memo. 834, 28 pp., https://www.ecmwf.int/en/elibrary/18748-use-situ-surface-observations-ecmwf.

Hamill, T. M., and S. Colucci, 1997: Verification of eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, 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.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for statistical post-processing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. *Quart. J. Roy. Meteor. Soc.*, **144**, 1257–1278, https://doi.org/10.1002/qj.3130.

Jolliffe, I. T., 2017: Probability forecasts with observation error: What should be forecast? *Meteor. Appl.*, **24**, 276–278, https://doi.org/10.1002/MET.1626.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Lopez, P., 2011: Direct 4D-Var assimilation of NCEP Stage IV radar and gauge precipitation data at ECMWF. *Mon. Wea. Rev.*, **139**, 2098–2116, https://doi.org/10.1175/2010MWR3565.1.

——, 2013: Experimental 4D-Var assimilation of SYNOP rain gauge data at ECMWF. *Mon. Wea. Rev.*, **141**, 1527–1544, https://doi.org/10.1175/MWR-D-12-00024.1.

——, G.-H. Ryu, B.-J. Sohn, L. Davies, C. Jakob, and P. Bauer, 2011: Specification of rain gauge representativity error for data assimilation. ECMWF Tech. Memo. 647, 24 pp., https://www.ecmwf.int/sites/default/files/elibrary/2011/10801-specification-rain-gauge-representativity-error-data-assimilation.pdf.

Massonnet, F., O. Bellprat, V. Guemas, and F. Doblas-Reyes, 2016: Using climate models to estimate the quality of global observational data sets. *Science*, **354**, 452–455, https://doi.org/10.1126/science.aaf6369.

Monache, L. D., J. Hacker, Y. Zhou, X. Deng, and R. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, https://doi.org/10.1029/2005JD006917.

Pillosu, F., and T. Hewson, 2017: New point-rainfall forecasts for flash flood prediction. *ECMWF Newsletter*, No. 153, ECMWF, Reading, United Kingdom, 4–5, https://www.ecmwf.int/en/newsletter/153/news/new-point-rainfall-forecasts-flash-flood-prediction.

Pinson, P., and R. Hagedorn, 2012: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteor. Appl.*, **19**, 484–500, https://doi.org/10.1002/met.283.

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.

Röpnack, A., A. Hense, C. Gebhardt, and D. Majewski, 2013: Bayesian model verification of NWP ensemble forecasts. *Mon. Wea. Rev.*, **141**, 375–387, https://doi.org/10.1175/MWR-D-11-00350.1.

Saetra, O., H. Hersbach, J.-R. Bidlot, and D. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501, https://doi.org/10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Santos, C., and A. Ghelli, 2012: Observational probability method to assess ensemble precipitation forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 209–221, https://doi.org/10.1002/qj.895.

Scheuerer, M., and T. 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.

Skamarock, W. C., J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and C-grid staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, https://doi.org/10.1175/MWR-D-11-00215.1.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106**, 11 775–11 784, https://doi.org/10.1029/2001JD900066.

Weber, N. J., and C. F. Mass, 2019: Subseasonal weather prediction in a global convection-permitting model. *Bull. Amer. Meteor. Soc.*, **100**, 1079–1089, https://doi.org/10.1175/BAMS-D-18-0210.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

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

——, 2019: Indices of rank histogram flatness and their sampling properties. *Mon. Wea. Rev.*, **147**, 763–769, https://doi.org/10.1175/MWR-D-18-0369.1.

——, and S. Vannitsem, 2018: Uncertain forecasts from deterministic dynamics. *Statistical Postprocessing of Ensemble Forecasts*, 1st ed. S. Vannitsem, D. Wilks, and J. Messner, Eds., Elsevier, 1–13.

Yamaguchi, M., S. T. K. Lang, M. Leutbecher, M. J. Rodwell, G. Radnoti, and N. Bormann, 2016: Observation-based evaluation of ensemble reliability. *Quart. J. Roy. Meteor. Soc.*, **142**, 506–514, https://doi.org/10.1002/qj.2675.