Technical Memo

European Centre for Medium-Range Weather Forecasts

865

Accounting for representativeness in the verification of ensemble forecasts

Zied Ben Bouallègue (Forecast Department)

Part of this work published in Monthly Weather Review

June 2020

Series: ECMWF technical memoranda

A full list of ECMWF Publications can be found on our web site under:
http:// www.ecmwf.int/ en/ publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

Spatial variability of 2 m temperature, 10 m wind speed, and daily precipitation is analysed to characterize to what extent measurements at a single location are representative of averages over a larger area. Characterization of representativeness error is made in probabilistic terms using parametric approaches, namely by fitting a normal, a truncated normal, and a censored shifted gamma distribution to observation measurements for the 3 weather variables of interest, respectively. Distribution parameters are estimated with the help of high-density network observational datasets. These results serve as a basis for accounting for representativeness error in ensemble verification. Uncertainty associated with the scale mismatch between forecast and observation is accounted for by applying a perturbed ensemble approach before the computation of scores. For all 3 variables investigated, verification results presented here quantify the large impact of representativeness error on forecast reliability and skill estimates.

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. More generally, observation errors in forecast verification have gathered more attention as the accuracy of the forecast in the short range approaches the accuracy of observation measurements (Saetra *et al.*, 2004; Candille and Talagrand, 2008; Pappenberger *et al.*, 2009; Santos and Ghelli, 2012; Röpnack *et al.*, 2013; Massonnet *et al.*, 2016; Jolliffe., 2017; Ferro, 2017; Duc and Saito, 2018).

It has been shown 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 which aim to capture the forecast uncertainty (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 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 over-spread, 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).

In order to account for observation uncertainty in the ensemble verification process, we have first to characterize observation errors. Observation errors are the sum of measurement errors and representativeness errors (RE). In the following, we focus on RE which is assumed to be the dominant contribution to observation errors associated with synoptic station (SYNOP) measurements in our applications. RE can be described in probabilistic terms as the relationship between observations at two different spatial scales. Statistical models can be used to describe the characteristics of representativeness for different weather variables. Such statistical methods have been developed in the context of ensemble post-processing to account for model limitations in representing sub-grid variability and correct simultaneously for systematic forecast deficiencies such as biases (Wilks and Vannitsem, 2018). Here, the parametric models are fitted exclusively with data from a high-density observation network. Thus, systematic model errors are not accounted for and can still be diagnosed in the verification.

The estimated uncertainty associated with station measurements is then used in the process that compares ensemble forecasts against SYNOP observations. We follow here the so-called perturbed ensemble method which consists in adding observation uncertainty to the forecasts. Perturbations drawn from appropriate parametric distributions are added to each ensemble member before computing probabilistic scores and diagnostic measures. The impact of accounting for observation representativeness on ensemble verification results is assessed and discussed for 2 m temperature, 10 m wind speed, and daily precipitation.

The main aim of this manuscript is to provide the reader with a generalized uncertainty model for these 3 weather variables. We present a fully parametric description of representativeness as a function of the aggregating scale (model grid resolution). Possible applications encompass not only ensemble forecast verification but also ensemble forecast post-processing. The expected impact on probabilistic forecast skill is documented in this Technical Memorandum. In Section 2, we introduce the data and the general methodology applied for model fitting, model validation, and forecast verification. In Section 3, we provide details about the models developed for each variable, the corresponding validation results, and ensemble scores showing the impact of accounting for representativeness. Conclusions and a discussion on future prospects can be found in Sections 4 and 5, respectively.

2 General methodology

2.1 Data

The analysis of RE for observation measurements relies on high-density observations (HDOBS). The data consists of observations provided by ECMWF Member and Co-operating States in addition to observations from the SYNOP stations available through the Global Telecommunication System (Haiden *et al.*, 2018). The observation network covers Europe and the station density relative to SYNOP is typically enhanced by a factor of 2-10.

The spatial coverage differs for each variable and from one day to another. Figures 1 shows the distribution of 2 m temperature, 10 m wind speed, and daily precipitation observations on 1 January 2018. Note that for precipitation, the focus is exclusively on an accumulation period of 24 h. The amount of RE will depend on the precipitation accumulation period, with less RE expected for longer accumulation periods than for shorter ones. Additionally, we can note that while wind speed is reported as an integer number in SYNOP messages, this is not the case for measurements from the HDOBS network.

2.2 Parametric models

For each variable, we propose a parametric model that aims to capture the variability on unrepresented scales. Inspired by existing ensemble post-processing methods, we consider the following probability distributions:

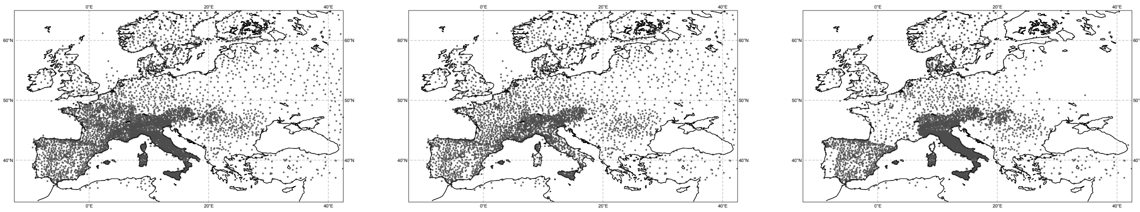

Figure 1: 2 m temperature (left), 10 m wind speed (middle), and 24 h precipitation (right) observation distribution from the HDOBS network on the 01/01/2018.

- a normal distribution for 2 m temperature,
- a truncated normal distribution for 10 m wind speed,
- a censored shifted gamma distribution for daily precipitation.

Each parametric distribution is fitted in the form of a conditional distribution for an observed quantity at one spatial scale, say B, given the same quantity aggregated over a larger scale, say A. More precisely, we are interested in the conditional probability:

$$P(Y_{\mathrm{B}} \mid Y_{\mathrm{A}}),$$

which is the probability of the random variable Y_{B}, representing the observation at a smaller scale, given the random variable Y_{A}, representing the observation at a larger scale (for example the grid scale of an NWP model). The aim is to characterize the relationship between averaged values over an area A and point measurements at B, where B is a point within the area A. This characterization will define the RE associated with point observations (such as SYNOP measurements) and used later in forecast verification.

2.3 Model fitting

Point measurements are compared with areal averages of observations. Neighbourhood areas A defined as square areas with length Δ_{A} and single observations (denoted y_{B}) within the areas are collected. When at least 5 observations are found within an area, the averaged quantity is computed and denoted y_{A}. Repeating this for all days of the data set, we obtain a sample of pairs $(y_{\mathrm{A}}, y_{\mathrm{B}})$. We note that there is an uncertainty associated with this method, as we do not know the actual value of y_{A}, just an estimate based on a limited set of point observations*.

Each parametric model is fitted with the pairs $(y_{\mathrm{A}}, y_{\mathrm{B}})$ in order to describe in probabilistic terms the relationship between these two quantities. The parameters are estimated for a range of neighbourhood sizes (Δ_{A}: 20, 30, ..., 140, 150 km). Eventually, the distribution parameters are described as a function of the size of the averaging area Δ_{A} using simple functions as described later.

The distribution parameters are estimated by minimizing the mean continuous ranked probability score (CRPS) over a test sample. Following Gneiting and Raftery (2007), the CRPS is defined for a distribution $F(y_{\mathrm{A}})$ and an observation y_{B} as follows:

$$\mathrm{CRPS} = E_X \mid X - y_{\mathrm{B}} \mid - \frac{1}{2} E_{X,X'} \mid X - X' \mid \tag{1}$$

*The minimum number of observations per grid box required to compute averaged precipitation is here set to 5. Increasing this number has little impact on the final results.

where X and X' are independent random variables drawn from the corresponding parametrized distribution. CRPS can be expressed in a closed form for the type of distributions used in this study (Gneiting *et al.*, 2008; Thorarinsdottir and Gneiting, 2010; Scheuerer and Hamill, 2015).

The learning sample for model fitting corresponds to 4 non-consecutive months of data (January, April, July, and October 2018). Model validation is performed on independent data-sets, namely samples corresponding to the months of February, May, August, and November 2018.

2.4 Model validation

The validity of the parametric models is checked by means of Probability Integral Transform (PIT, Raftery *et al.*, 2005) histograms. The following diagnostic procedure is applied: we consider percentiles associated with the parametric distributions 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 similarly as rank histograms (Hamill and Colucci, 1997), where a histogram close to a uniform distribution indicates reliability.

In addition, PIT histograms are generated separately for sub-samples of the validation data-sets. Stratification is based on the value of the area averaged quantity y_{A}. Stratified PIT histograms are produced for equi-populated categories (using terciles) corresponding to cases with low, intermediate, and high y_{A} values. Stratified PIT histograms help diagnose potential limitations of the parametric models.

Finally, we perform a visual inspection of Quantile-Quantile (Q-Q) plots for random draws of the parametric distribution and a set of points observations. Q-Q plots help diagnosing whether the two sets (model draw and point observation) are drawn from the same marginal distribution.

2.5 Perturbed ensemble approach

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

Practically, random noise is 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 centred over the forecast value and its spread accounts for representativeness uncertainty. This approach can also be seen as a forecast down-scaling that provides a description of the sub-grid scale uncertainty which 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.

2.6 Verification data-sets

The performance of forecasts from ECMWF's medium-range ensemble forecasts (ENS) is assessed with and without accounting for RE. In the ensemble verification process, only SYNOP measurements are used. They are compared to the corresponding nearest grid-point forecasts. The horizontal grid spacing of ENS forecasts is about 18 km for the main verification period considered here (JJA 2018)[†].

Results for JJA 2018 are based on forecasts initialised at 12 UTC with a verification domain corresponding to the European continent (see Figure 1). In addition, the proposed models are tested over a winter period (DJF 2019) but the corresponding results are partially shown in the Appendix and only discussed in the text when appropriate.

Furthermore, analysis of long-term trends in forecast performance is carried out. For this, we use forecasts initialised at 12 UTC over the past 16 years (September 2004 to September 2019). Over that period, the grid resolution of the ensemble forecasts has evolved as follows: 80 km until February 2006, 60 km until January 2010, 35 km until March 2016, and 18 km since then. These changes in grid-resolution have to be accounted for in order to adapt accordingly the parameters of the distributions used in the perturbed ensemble approach.

2.7 Verification metrics

First, we compute a general measure of ensemble performance for continuous variables, namely the CRPS (see Eq. 1). Second, the impact of accounting for RE in the verification process is assessed by focusing on binary events. Summary performance measures for probability forecasts are computed: we apply the Brier score (BS, Brier, 1950) and the diagonal elementary score (DES, Ben Bouallègue *et al.*, 2018) with the verification sample climatology used for the computation of the event base-rate. Block boot-strapping with blocks of 3 days and 1000 iterations is used to estimate confidence intervals.

Additionally, 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 non-event. In this study, the sharpness diagram, which is usually included in reliability diagrams, is plotted separately. 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.

Finally, we assess the impact of accounting for representativeness on long-term trends of ensemble forecast performance focusing on extreme events. We follow the methodology recommended in Ben Bouallègue *et al.* (2019). For each variable, DES in the form of a skill score is estimated for a threshold defined as the 95% percentile of the climate distribution. Different climate distributions are considered for the forecast and the observations, respectively. This type of threshold definition, referred to as the eigenclimatology approach, allows to eliminate a potential mismatch between forecast and observation marginal distributions at the station level.

[†] Note that the verification task is independent of the representativeness model fitting and model validation taks.

[‡] Numerically, ROC curves are generated using the 51 possible probability thresholds issued by the 50-member ensemble.

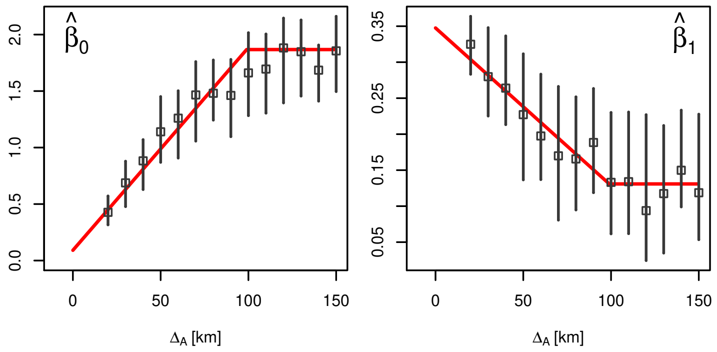

Figure 2: Estimated parameters for the Normal model, $\hat{\beta}_0$ and $\hat{\beta}_1$, as a function of the averaging scale Δ_A (in km) with 5%-95% confidence intervals estimated with block-bootstrapping. In red, the generalized model uncertainty model for 2 m temperature observations.

3 Results

3.1 Temperature at 2 m above ground

For 2 m temperature, RE is already (partially) accounted for routinely in forecast verification. A standard procedure consists in a bias correction, with an adjustment linear with the height difference between station and model grid:

$$t^\star = t + 0.0065 D_e \quad \text{with} \quad D_e = e_\mathrm{m} - e_\mathrm{o} \quad , \tag{2}$$

with $t^\star$ and t the adjusted and raw forecasts, respectively, and where e_o and e_m denote the elevation of the observation and the corresponding model elevation at the nearest model grid point, respectively.

Here, 2 m temperature representativeness uncertainty is described in probabilistic terms with the help of a normal distribution $\mathcal{N}(\mu,\sigma)$ with mean μ and standard deviation σ. After exploratory analysis of the spatial variability of 2 m temperature observations, we propose the following model:

$$\mu = t^\star, \quad \sigma = \beta_0 + \beta_1 \sqrt[4]{\mid D_e \mid} \quad . \tag{3}$$

The standard adjustment procedure is followed regarding the mean of the distribution while the standard deviation σ comprises a constant term and a second term linear with a power transformation of $\mid D_e \mid$. This later aims to capture the increase of representativeness uncertainty with complex topography. The use of a $\frac{1}{4}$ power transformation is justified by the fact that it leads to better fit with point observations than the simple use of the elevation absolute difference or its square root. The optimization process is initialized with the following set of parameters: $(\beta_0 = 1, \beta_1 = 0.1)$.

The model parameters of (3) are estimated for a range of aggregating scales as shown in Figure 2. $\hat{\beta}_0$ exhibits a linear growth with the aggregating scale Δ_A before reaching a plateau and $\hat{\beta}_1$ decreases with Δ_A before reaching a plateau. By fitting a linear functions to describe $\hat{\beta}_0$ and $\hat{\beta}_1$ as a function of Δ_A up to $\Delta_\mathrm{A} = 100\,km$ and considering constant parameters for higher values of Δ_A, we eventually obtain the generalized uncertainty model for 2 m temperature observations:

$$\beta_0(\Delta_\mathrm{A}) = \begin{cases} 0.02\Delta_\mathrm{A} & \text{if} \quad \Delta_\mathrm{A} < 100\,km \\ 2 & \text{if} \quad \Delta_\mathrm{A} > 100\,km \end{cases} \qquad \beta_1(\Delta_\mathrm{A}) = \begin{cases} 0.35 - 0.002\Delta_\mathrm{A} & \text{if} \quad \Delta_\mathrm{A} < 100\,km \\ 0.15 & \text{if} \quad \Delta_\mathrm{A} > 100\,km \end{cases} \tag{4}$$

to be used in combination with (3).

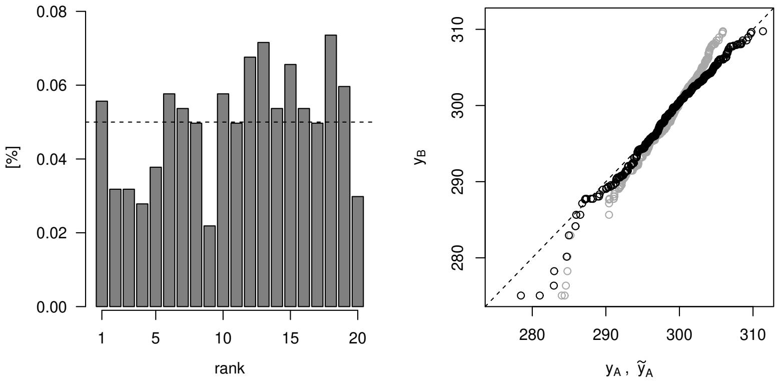

Figure 3: Model validation through PIT histograms (left), and Q-Q plot for the original sample y_A displayed in grey, and the fitted model $\tilde{y}_A$ displayed in black (right) expressed in K. Results for an averaging scale Δ_A of 40 km, August 2018.

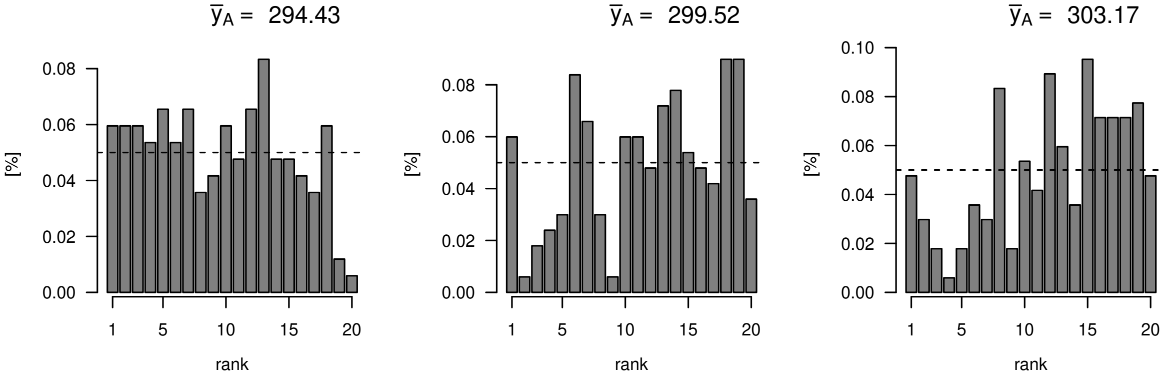

Figure 4: Model validation through stratified PIT histograms for the same data-set as in Figure 3. Results for equi-populated categories corresponding to low, intermediate, and high averaged temperatures y_A are shown form left to right, respectively. The corresponding mean y_A in K is indicated above each histogram.

From Figure 3, we see that this model seems overall appropriate. The PIT histogram on the left panel is rather noisy but with no large departure from uniformity and the Q-Q plot on the right panel indicates that the model captures well the empirical marginal distribution of the point observations (except for extreme low temperature). The inspection of stratified PIT histograms in Figure 4 reveals an asymmetry in the rank distribution between low and high temperature categories. The distribution of point observations seems not perfectly captured with random draws from the model: point observations tend to be slightly overestimated (underestimated) in case of low (high) area average temperatures. This underestimation of the RE with respect to potential low and high temperature occurs in Spring and Summer, while in Winter RE is overestimated in cold conditions (not shown). Whereas the model reliability is overall reasonably good, RE modelling would certainly benefit from accounting for RE diversity as a function of the weather situation.

Now the model presented in (4) is used to account for RE in the verification process of ENS 2 m temperature forecasts. The averaging scale Δ_A corresponds now to the horizontal grid spacing of ENS. The set of β parameters for an averaging scale of 18 *km* is approx. (0.4,0.3). This set of parameters is used in the perturbed ensemble approach and results are shown in Figures 5 to 7.

In Figure 5, we see on the right panel that accounting for RE leads to an improvement in terms of CRPSS of about 11% at day 1 (8% in Winter, Fig. 23 in the Appendix). The impact of the perturbed ensemble approach is fading with the forecast horizon but is still non negligible at day 15. In Winter, the impact drops below 1% for forecast lead times greater than 7 days. In absolute term, the CRPS plotted on the left panel increases with lead time from day 1 when applying the perturbed ensemble approach. This is not visible for the original results (grey line) but this feature is consistent with the expected error growth of the forecast.

In Figure 6, the focus is on the ensemble performance at day 5. Brier skill score (BSS, Figure 6.a) and diagonal elementary skill score (DESS, Figure 6.b) compare skill before and after applying the perturbed ensemble approach. BSS and DESS are plotted here for a range of event-thresholds. The impact on BS varies between 3 and 7% while the impact on DES can become large (over 30%) for low and high temperature thresholds. Verification over the Winter indicates an improvement of around 1% in terms of BS and 10 % in terms of DES for an event defined as 2 m temperature dropping below 0 $^{\circ}C$ (Fig. 24 in the Appendix). The difference between BS and DES results can be related to the impact of RE on ensemble forecast attributes which is investigated in more detail now.

In Figure 7, the focus is on the ensemble forecast attributes for a threshold of 293 K at day 5: reliability (left panel), sharpness (middle panel), and discrimination (right panel). An increase in reliability, a decrease in sharpness, and a slight improvement of the discrimination ability are visible when accounting for RE[§]. The ensemble derived probabilities for this event appear close to reliability when accounting for RE. This is not the case over the winter period where underdispersivness is still visible even after applying the perturbed ensemble approach (Fig. 25 in the Appendix). The impact on the forecast reliability predominantly explains the improvement seen in BS. It comes at the cost of a decrease in forecast sharpness. In terms of discrimination, the impact appears small in Figure 7 but is larger for more rare events (Fig. 26 in the Appendix). This is consistent with the impact seen in terms of DES. This increase in discrimination ability for rare observed events is explained by the ability of the perturbed ensemble to exceed these event-thresholds more often, and so to distinguish better between low probability forecasts.

Finally, Figure 8 compares long-term trends in forecast performance focusing on extreme events before (left panel) and after (right panel) applying the perturbed ensemble approach. DESS for an event defined as the 95% percentile of the eigenclimatology is used to measure the forecast performance. Qualitatively, the trends are very similar for the 2 plots, but we see a clear differentiation between performance at different lead times after applying the perturbed ensemble approach. Results for day 1 and day 2 are the most impacted which is consistent with the large impact of representativeness on forecast performance at short lead time seen in Figure 5. Quantitatively, forecast skill can appear smaller after accounting for RE. This is the case for longer lead times and lower grid-resolution forecasts (i.e. before 2010). At low-resolution, the perturbed ensemble approach adds a level of noise that could hinder to capture the forecast signal at longer lead times when the signal-to-noise ratio in the forecast is small. Figure 8 also shows a more monotonous increase in forecast skill over time when RE is accounted for, as well as a more consistent dependence on lead time.

[§] Note that the same number of points (same probability levels) is used for all ROC curves.

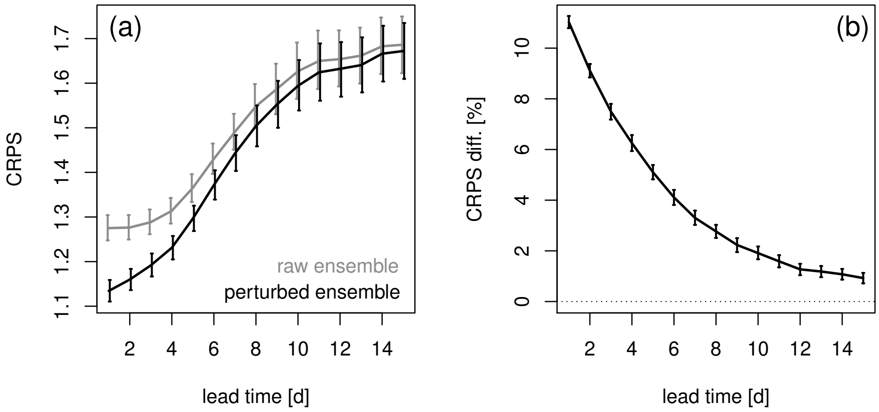

Figure 5: (a) CRPS (in K) estimated with (black) and without (grey) accounting for representativeness uncertainty and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Vertical bars indicate 95% confidence intervals. Results valid for 2m temperature ensemble forecasts, Europe, Summer 2018.

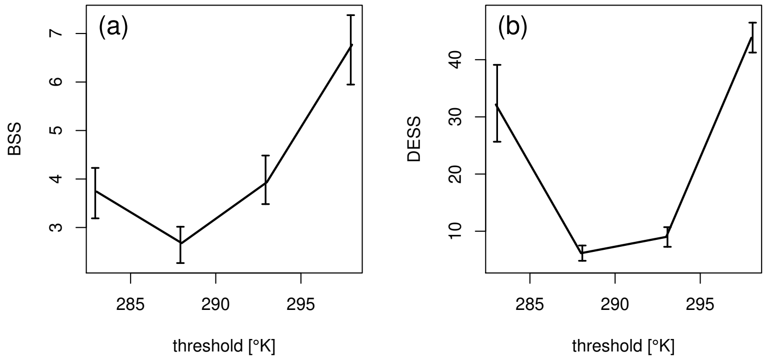

Figure 6: (a) Brier skill score (BSS), and (b) diagonal elementary skill score (DESS) as a function of the event-threshold when verifying 2m temperature ensemble forecasts at day 5. Vertical bars indicate 5%-95% confidence intervals. Results valid for Europe, Summer 2018.

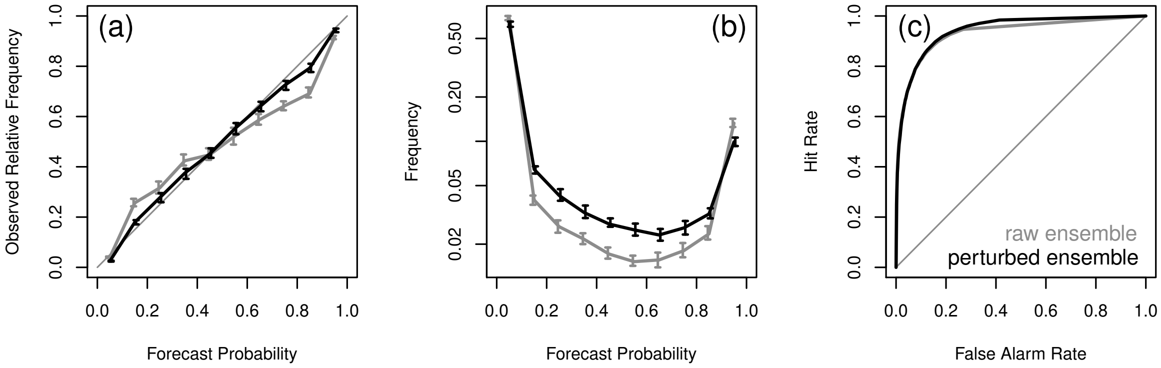

Figure 7: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of 293 K. Results with (black) and without (grey) accounting for representativeness uncertainty when verifying ensemble forecasts at day 5. Vertical bars indicate 95% confidence intervals. Results valid for 2m temperature ensemble forecast, Europe, Summer 2018.

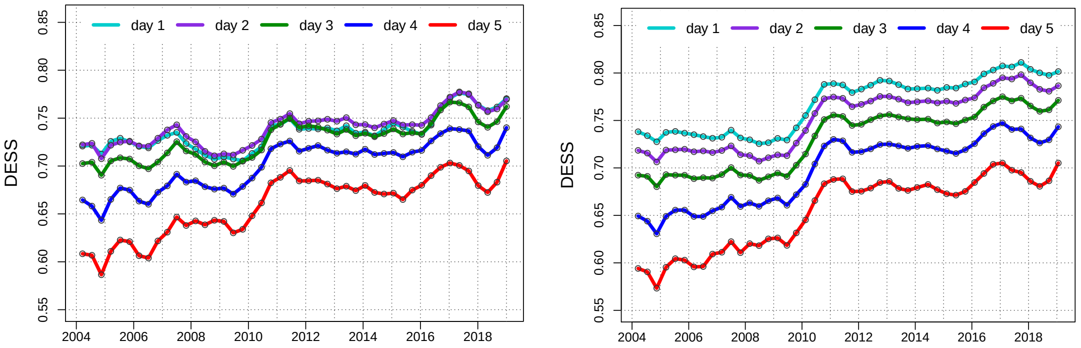

Figure 8: Long-term trends in 2m temperature ensemble forecast performance as measured by the diagonal elementary skill score (DESS) with (right) and without (left) accounting for representativeness uncertainty. 1-year running mean over the period Sept. 2004 - Sept. 2019. The event under focus is defined as the 95% percentile of the eigenclimatology.

3.2 Wind speed at 10 m above ground

Representativeness associated with 10 m wind speed observations is described with the help of a truncated normal distribution with cut-off at zero $\mathcal{N}^0(\mu,\sigma^2)$ with location μ and shape σ parameters. The probability distribution function (PDF) follows:

$$\frac{\frac{1}{\sigma}\phi((x-\mu)/\sigma)}{\Phi(\mu/\sigma)}, \quad x \geq 0, \quad \text{and} \quad 0 \quad \text{otherwise}, \tag{5}$$

with ϕ and Φ the PDF and cumulative distribution function (CDF) of the standard normal distribution, respectively. This type of distribution was successfully used for the post-processing of 10 m wind speed ensemble forecasts (Thorarinsdottir and Gneiting, 2010).

Here, we propose a parametric model based on 3 parameters $(\alpha_0, \alpha_1, \beta_1)$ in order to describe the location and shape of the distribution of point measurements as a function of the area averaged quantity y_B:

$$\mu(y_B) = \alpha_0 + \alpha_1 y_B \quad \text{and} \quad \sigma(y_B) = \beta_1\sqrt{y_B} + \varepsilon \tag{6}$$

where ε is set to 0.01 in order to avoid a possible division by zero in (5). The distribution location μ is adjusted by means of the intercept α_0 and the multiplicative parameter α_1. The shape of the distribution is a function of the square root of the area averaged wind speed with multiplicative factor β_1. The optimization process is initialized with the following set of parameters: $(\alpha_0 = 0, \alpha_1 = 1, \beta_1 = 0.5)$.

The parameters of the model in (6) estimated for a range of averaging scales are shown in Figure 9. $\hat{\alpha}_0$ appears linear with Δ_{A} and tends to 0 when Δ_{A} decreases. $\hat{\alpha}_1$ also appears linear with Δ_{A} but tends to 1 as Δ_{A} decreases. $\hat{\beta}_1$ converges to 0 as the averaging scale decreases and has a parabolic shape when plotted as a function of Δ_{A}. Fitting the appropriate function for each parameter, we obtain the following generalized uncertainty model for 10 m wind speed:

$$\alpha_0(\Delta_{\text{A}}) = -0.02\Delta_{\text{A}}, \quad \alpha_1(\Delta_{\text{A}}) = 1 + 0.002\Delta_{\text{A}}, \quad \beta_1(\Delta_{\text{A}}) = -0.04\Delta_{\text{A}} + 0.17\Delta_{\text{A}}^{0.75}. \tag{7}$$

to be combined with the truncated normal distribution defined in (6).

Validation of the model in (7) is shown in Figure 10. A very good agreement is visible in terms of Q-Q plot being close to the diagonal, but the PIT histogram indicates a slight overestimation of the

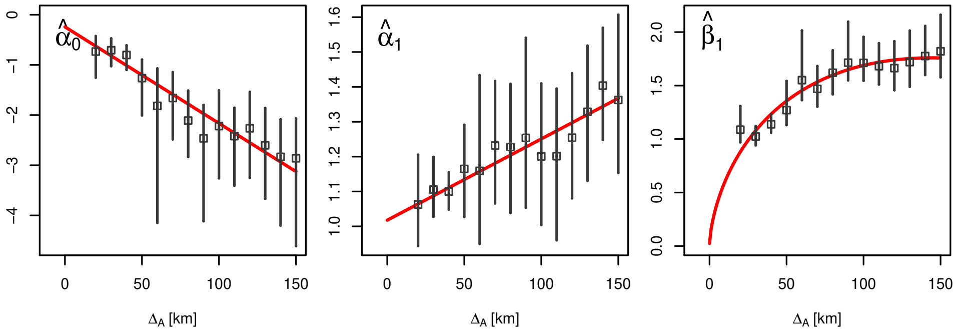

Figure 9: Estimated parameters for the Truncated Normal model, $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\beta}_1$, as a function of the averaging scale Δ_A (in km) with 5%-95% confidence intervals estimated with block-bootstrapping. In red, the general uncertainty model for 10 m wind speed.

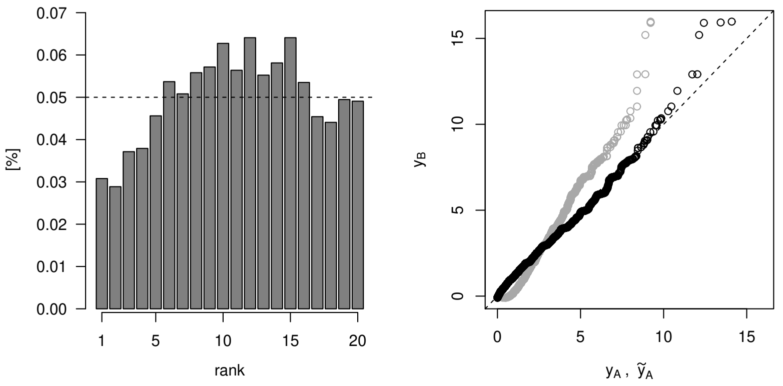

Figure 10: Model validation through PIT histograms (left), and Q-Q plot for the original sample y_A displayed in grey, and the fitted model $\tilde{y}_A$ displayed in black (right). Results for averaged observations with $\Delta_A = 40km$, August 2018.

uncertainty for this validation month (August). In Figure 11, stratified PIT histograms show that the RE overestimation is more pronounced for high wind categories while the model is positively biased for low and intermediate wind situations. The analysis of the rank histograms for the other validation months reveals that 10 m wind representativeness uncertainty is underestimated by the model in Autumn and Winter but is well captured in Spring (not shown). In particular, in Winter, the extreme right bin of the stratified rank histograms is overpopulated for the high wind speed category. This feature reflects that the parametric model underestimates the potential occurrence of very high wind speed at point location based on the average wind speed in the corresponding area. With an underestimation of RE in Winter and an overestimation in Summer, future work should focus on including seasonality in the RE model for 10 m wind speed.

Based on the generalized model in (7), we apply the following set of parameters for the perturbation of ENS 10 m wind speed forecasts over the Summer 2018: $(\alpha_0 = -0.36, \alpha_1 = 1.036, \beta_1 = 0.76)$. The corresponding results are shown in Figures 12 to 14.

Figure 12 shows the impact of representativeness on verification scores as a function of the forecast lead time. Both CRPS and CRPS relative differences are shown. Focusing on the latter one, we see a radical change in the shape of the curve. In a standard framework (without accounting for RE), the ensemble skill at day 1 is similar to the ensemble skill at day 15. When accounting for representativeness, CRPS

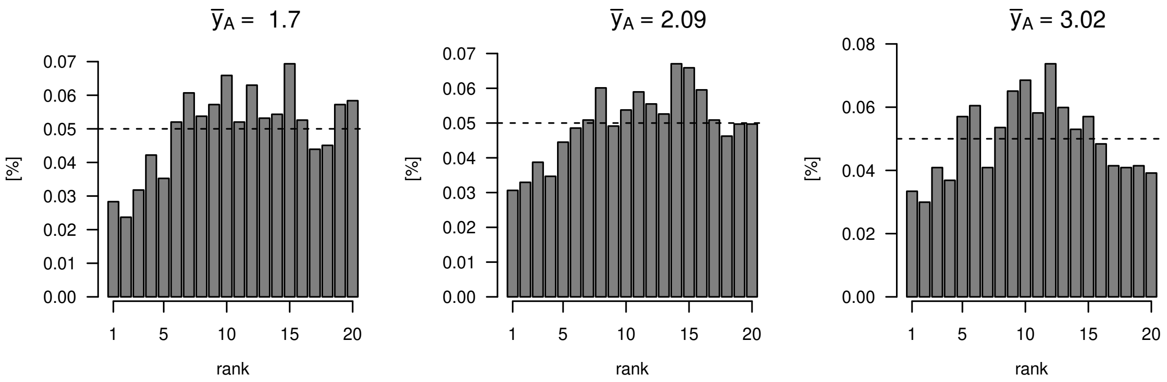

Figure 11: Model validation through stratified PIT histograms for the same data-set as in Figure 10. Results for equi-populated categories corresponding to low, intermediate, and high averaged wind speed y_A are shown form left to right, respectively. Mean y_A in m/s for each category is indicated on the top of the histograms.

is increasing as a function of the forecast lead time up to around day 7 where the forecast skill reaches a plateau. In relative terms, the impact of RE is large in particular at short lead time, about 15% at day 1, and decays to a around 6% at day 7. Over the winter period, the corresponding results are 14% at day 1 and about 4% after day 10 (Fig. 27 in the Appendix).

Focusing on day 5, Figure 13 summarises the RE impact on the ensemble verification results in terms of BSS and DESS plotted as a function of the event-threshold. On one side, in Figure 13.a, BSS reaches 2% for a threshold of 5 m/s while significant improvements are not visible for larger thresholds. Because the CRPS corresponds to the integral of the BS over all possible thresholds, the large impact in terms of CRPS can be attributed mostly to skill differences for small threshold events. On the other side, in Figure 13.b, DESS can exceed 10% for intermediate thresholds (10 or 15 m/s). Similar results are found over the winter period, with DESS close to 10% not only for intermediate but also for larger thresholds.

Considering an event-threshold of 5 m/s, Figure 14 depicts the impact of representativeness on forecast reliability, forecast probability distribution, and forecast discrimination. Results are valid for a forecast lead time of 5 days. As expected, a large impact on forecast reliability and forecast sharpness is visible: the reliability curve is closer to the diagonal indicating good reliability and, at the same time, low (high) forecast probabilities are more (less) frequent which indicates less sharp forecasts. There is no visible impact on forecast discrimination for this threshold. Considering higher thresholds, reliability curves become noisier and difficult to interpret while ROC curves are superimposed but with extra points along the virtual full ROC curve when accounting for representativeness (Fig. 28 in the Appendix).

Finally, Figure 15 focuses on long-term trends in performance. Considering extreme events defined as the 95% of the eigenclimate distribution, annual running mean of DESS are compared before (left panel) and after (right panel) applying the perturbed ensemble approach. Dissimilarities between the two plots are more important for short lead times, performance trends at day 1 and day 2 in particular. For both plots, the eigenclimatology approach corrects for potential differences between forecasts and observations climate distributions. Additionally, the perturbed ensemble approach for 10m wind speed (right panel) increases the spread as a function of the forecast wind speed intensity. This could explain the large improvement in terms of DESS for short lead times (day 1 and 2) when accounting for RE. As for 2m temperature, improvements over time due to improvements in the forecasting system are more clearly visible when RE is taken into account.

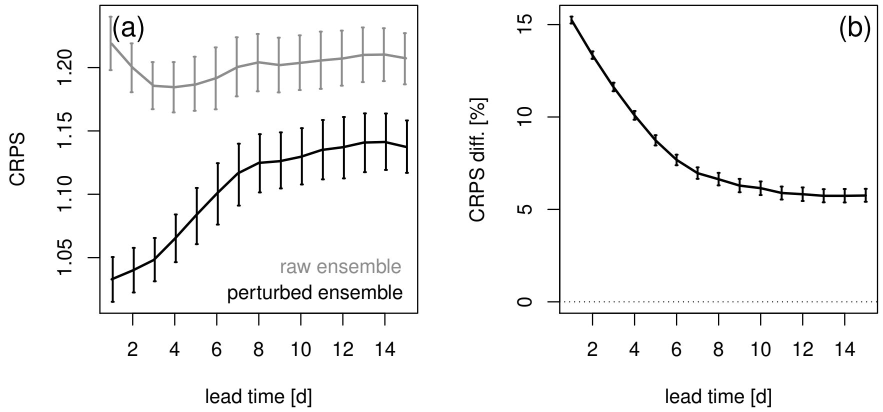

Figure 12: (a) CRPS (in m/s) computed with (black) and without (grey) accounting for representativeness uncertainty and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Vertical bars indicate 95% confidence intervals. Results valid for 10 m wind speed ensemble forecasts, Europe, Summer 2018.

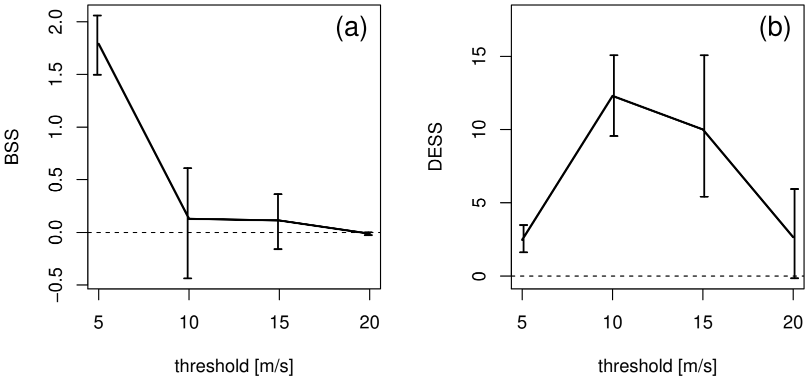

Figure 13: (a) Brier skill score (BSS), (b) and diagonal elementary skill score (DESS, right) comparing forecast performance before and after post-processing. Skill scores are shown as a function of the event-thresholds. Results are valid for 10 m wind speed forecasts at day 5, Europe, Summer 2018. Vertical bars indicate 5%-95% confidence intervals.

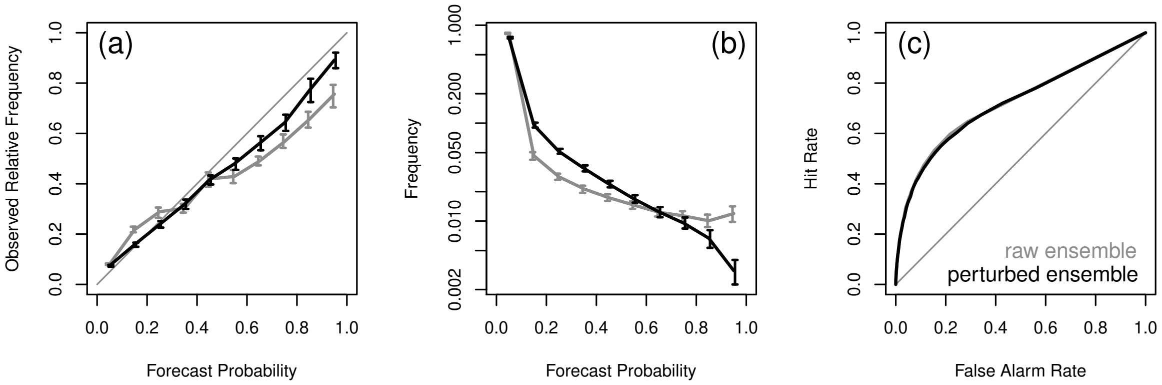

Figure 14: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of 5 m/s at day 5. Results with (black) and without (grey) accounting for representativeness uncertainty in the ensemble verification process. Vertical bars indicate 95% confidence intervals. Results valid for 10 m wind speed ensemble forecasts, Europe, Summer 2018.

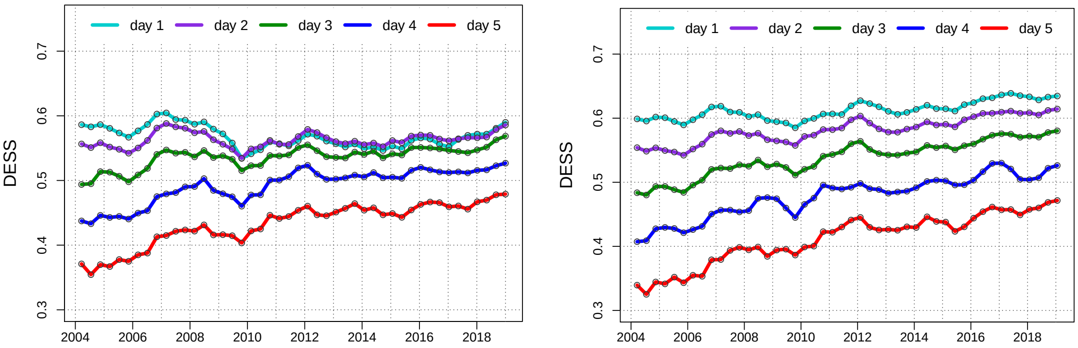

Figure 15: Long-term trends in 10 m wind speed ensemble forecast performance as measured by the diagonal elementary score (DESS) with (right) and without (left) accounting for representativeness uncertainty. 1-year running mean over the past 16 years. The event under focus is defined as the 95% percentile of the eigenclimatology.

3.3 Daily precipitation

The parametric model for daily precipitation representativeness is based on a censored, shifted gamma distribution (CSGD). This distribution has been successfully used in the post-processing of ensemble precipitation forecasts (Scheuerer and Hamill, 2015). This model appears therefore well-suited to describe precipitation RE and its peculiar characteristics: a probability distribution with a long tail and an uncertainty that grows with precipitation intensity.

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 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:

$$k = \frac{\mu^2}{\sigma^2} \quad \text{and} \quad \theta = \frac{\sigma^2}{\mu}. \tag{8}$$

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

$$\widetilde{F}_{k,\theta,\delta}(y) = \begin{cases} F_k\left(\frac{y+\delta}{\theta}\right) & \text{for} \quad y \geq 0 \\ 0 & \text{for} \quad y < 0 \end{cases} \tag{9}$$

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).

Exploratory analysis of the model sensitivity to the number of parameters suggests that 5 coefficients are required in order to describe the distribution of point observations appropriately. Two coefficients (α_0 and α_1) are associated with the mean of the distribution, μ_B:

$$\mu_\mathrm{B}(y_\mathrm{A}) = \alpha_0 + \alpha_1 y_\mathrm{A}, \tag{10}$$

which is a function of the averaged observed precipitation over an area A (y_A). Two other coefficients (β_0 and β_1) are associated with the standard deviation of the distribution, σ_B:

$$\sigma_\mathrm{B}(y_\mathrm{A}) = \beta_0 + \beta_1 \sqrt{y_\mathrm{A}}, \tag{11}$$

which is a function of the square root of the area averaged observed precipitation ($\sqrt{y_\mathrm{A}}$). The use of a power transformation in the relationship between precipitation intensity and uncertainty can be traced back to pioneering work on post-processing of ensemble precipitation forecasts (Hamill *et al.*, 2008). The fifth coefficient corresponds to δ which defines the shift associated with the CSGD. Optimization is performed using squared parameters to ensure that they are positive, and with the set ($\alpha_0 = 0.1$, $\alpha_1 = 1$, $\beta_0 = 0.1$, $\beta_1 = 1$ and $\delta = 0.1$) as initial values of the optimization process.

In Figure 16, 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 additive parameter $\hat{\alpha}_0$ is linearly increasing with Δ_A while the multiplicative parameter $\hat{\alpha}_1$ is constant (around 1) for all averaging scales. The shift parameter $\hat{\delta}$ is also increasing linearly with Δ_A. With similar values for $\hat{\alpha}_0$ and$\hat{\delta}$, the mean of the CSGD is close to y_A, which means that the expected mean precipitation intensity does not vary across scales.

One of the two coefficients associated with the variance of the distribution ($\hat{\beta}_0$) exhibits a slight increase with Δ_A while the other one ($\hat{\beta}_1$) has a large variability as a function of the averaging scale Δ_A. $\hat{\beta}_0$ and $\hat{\beta}_1$ influence the uncertainty associated with the CSGD distribution. Indeed, they determine the

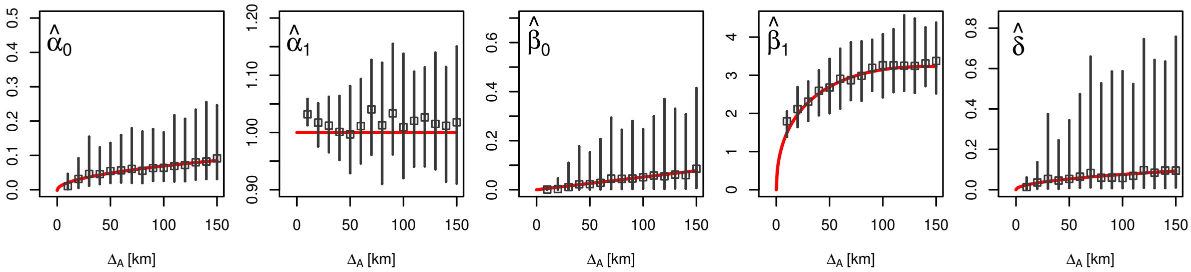

Figure 16: Estimated parameters for the Censored Shifted Gamma model, $\hat{\alpha}_o$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\delta}$, as a function of the averaging scale Δ_A (in km) with 25%-75% confidence intervals estimated with block-bootstrapping. The red lines represent the generalized uncertainty model for daily precipitation.

variance and skewness of the distribution through the shape parameter θ (Eq. 8). 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, RE of a single observation increases with the size of the target grid box, in agreement with results from previous studies (Lopez *et al.*, 2011).

Based on these results, we propose the following generalized uncertainty model for daily precipitation:

$$\alpha_0(\Delta_A) = 0.005\sqrt{\Delta_A}, \quad \alpha_1(\Delta_A) = 1,$$

$$\beta_0(\Delta_A) = 0.0005\Delta_A, \quad \beta_1(\Delta_A) = -0.02\Delta_A + 0.55\sqrt{\Delta_A},$$

$$\delta(\Delta_A) = 0.005\sqrt{\Delta_A} \tag{12}$$

to be combined with Eqs (10) and (11).

In Figure 17, this model is validated by means of PIT histogram and Q-Q plot. The histogram looks slightly U-shaped which indicates a slight underestimation of the representativeness uncertainty by the model. The Q-Q plot shows that the fitted model captures well the tail of the point observation distribution (black points). In other words, the CSGD approach allows to generate large precipitation amounts at an appropriate frequency.

The estimated parameters characterize precipitation RE throughout the year. However, seasonality in the magnitude of RE should be expected (Lopez *et al.*, 2011). Inspecting stratified rank histograms for August in Figure 18, we see that the model captures well small scale variability associated with low box-averaged precipitation but tends to underestimate precipitation variability associated with more intense box-averaged precipitation. Conversely, over winter months, the parametric model tends to overestimate spatial variability associated with large amount of area-averaged precipitation (not shown). This illustrates the limitation of the simplistic approach followed here. Parameters that vary as a function of the time of the year would be needed in order to tackle this deficiency. Future refinement of the present method could also consider, for example, CSGD parameters that vary as a function of weather situation (such as, for example, convective versus non-convective situation), orography, or region (*e.g.* tropics vs extratropics).

In Figure 19, the general impact of accounting for RE is shown: CRPS and relative CRPS difference are plotted as a function of the forecast lead time. Results with and without representativeness uncertainty are compared. A large impact is visible in particular at short lead times: from 12% at day 1, the relative difference becomes smaller than 2% after day 7 (left panel). Verification results for the Winter period are very similar (Fig. 29 in the Appendix). Since the ensemble spread (and forecast error) is small at the

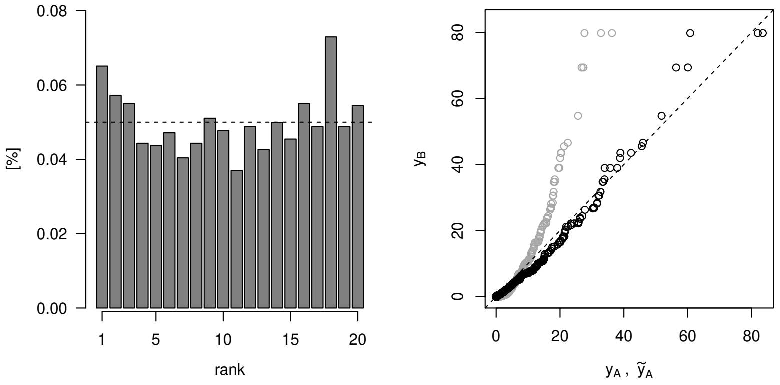

Figure 17: Model validation through PIT histograms (left), and Q-Q plot for the original sample y_A displayed in grey, and the fitted model $\tilde{y}_A$ displayed in black (right). Results for averaged observation for $\Delta_A = 40\,km$, August 2018.

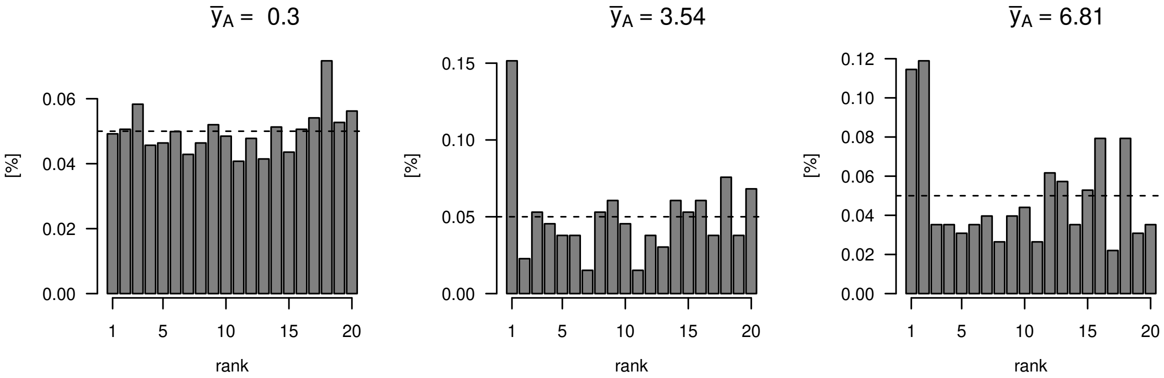

Figure 18: Model validation through stratified PIT histograms for the same data-set as in Figure 17. Results are shown for equi-populated categories corresponding to low, intermediate, and high y_A form left to right. Mean y_A in K for each category is indicated on the top of each histogram.

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.

The differences in terms of skill as measured by CRPS can be explained by the large improvement of the forecast reliability when accounting for RE. 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. In order to illustrate this point, reliability curves and sharpness diagrams for an event-threshold of 1 mm/24h are shown in Figures 20.a and 20.b. With representativeness uncertainty, the reliability curve is closer to the diagonal while the sharpness of the ensemble is noticeably decreased. For higher thresholds (Fig. 30 in the Appendix), the lack of reliability of the ensemble even after addressing representativeness provides evidence that there are remaining system problems that likely need to be addressed through prediction system improvement and/or post-processing.

Now, we inspect the role of RE when assessing ensemble forecasts in terms of discrimination. The right panels in Figure 20 show the impact of accounting for observation errors on ROC curves. For a 1 mm/24h threshold-event, the impact is neutral: the two curves that are compared are on top of each other. The information content of the forecast is not modified for this type of event when adding the representativeness uncertainty to the forecast. However, when focusing on larger event-thresholds, such as 20 mm/24h, the area under the two curves clearly differs in terms of extent (Fig. 30 in the Appendix). 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 which appears beneficial for users with small probability thresholds. The ability of the perturbed ensemble to forecast large values, and so to capture the tail of the observation distribution, is rewarded in terms of forecast discrimination. However, there is no real increase in terms of information content since no shift of the ROC curve towards the top left corner of the diagram is registered.

Figure 21 provides a summary of the forecast performance at day 5 as a function of the event threshold. In terms of BSS, large impact is noted for low-intensity events while in terms of DESS, larger differences are visible for more high-intensity events. This result is consistent with general characteristics of BS and DES: BS is more sensitive to reliability while DES is more sensitive to discrimination (Ben Bouallègue *et al.*, 2019).

Finally, trends in forecast performance with and without accounting for RE are compared in Figure 22. The main difference between the two plots consists in higher DESS at day 1 and slightly lower DESS at longer lead time when accounting for representativeness. In particular, performance at day 1 are distinguishable from performance at day 2 on the right panel. Overall, the trends for the different lead times are qualitatively very similar before and after applying the perturbed ensemble. The quantitative differences (and their interpretation) are in line with the ones seen in 2 m temperature and 10 m wind speed long-term performance.

4 Conclusion

This report provides a general methodology for accounting for representativeness when verifying ensemble forecasts. First, parametric models, based on normal, truncated normal, and censored shifted gamma distributions, are fitted with high density observations in order to describe the representativeness uncertainty associated with 2 m temperature, 10 m wind speed, and precipitation station measurements, respectively. These models are successfully validated by means of PIT histograms and Q-Q plots, but limitations of each model are also pointed out. Second, a perturbed ensemble approach is applied: it consists in perturbing each ensemble member by means of the proposed parametric models. This step

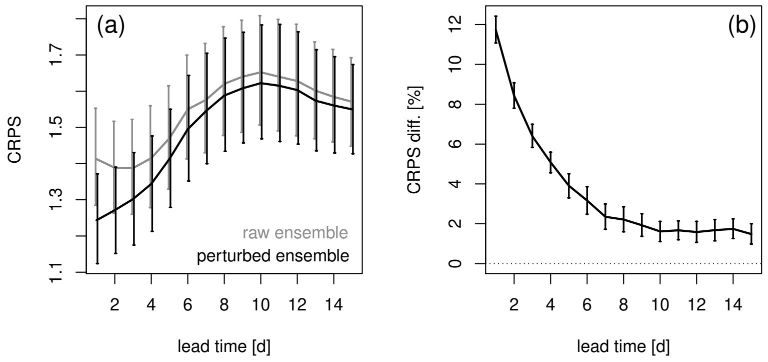

Figure 19: (a) CRPS (in mm/24h) with (black) and without (grey) accounting for representativeness uncertainty, and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Vertical bars indicate 95% confidence intervals. Results valid for daily precipitation ensemble forecasts, Europe, Summer 2018. (Published as Fig. 7 in Ben Bouallègue et al. (2020). © American Meteorological Society. Used with permission.)

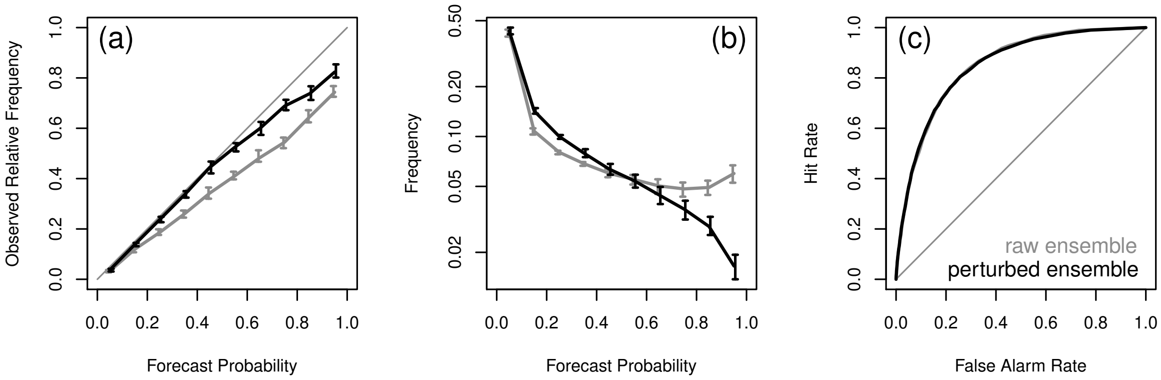

Figure 20: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of 1 mm/24h. Results with (black) and without (grey) accounting for representativeness uncertainty when verifying ensemble precipitation forecasts at day 5. Vertical bars indicate 95% confidence intervals. Results valid for daily precipitation ensemble forecasts, Europe, Summer 2018. (Published as Fig. 9 in Ben Bouallègue et al. (2020). © American Meteorological Society. Used with permission.)

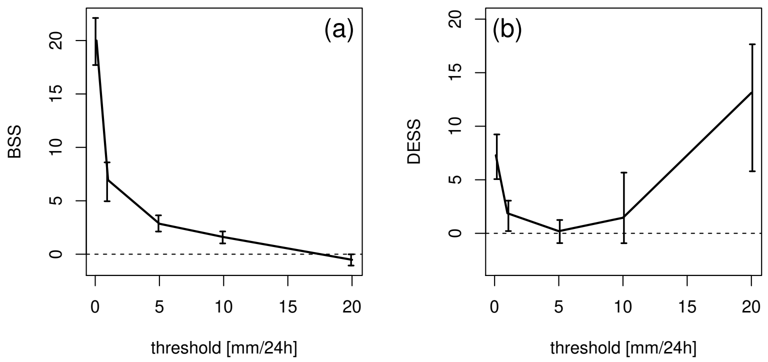

Figure 21: (a) Brier skill score (BS), and (b) diagonal elementary skill score (DESS) as a function of the event-thresholds when verifying daily precipitation forecast at day 5. Vertical bars indicate 5%-95% confidence intervals.

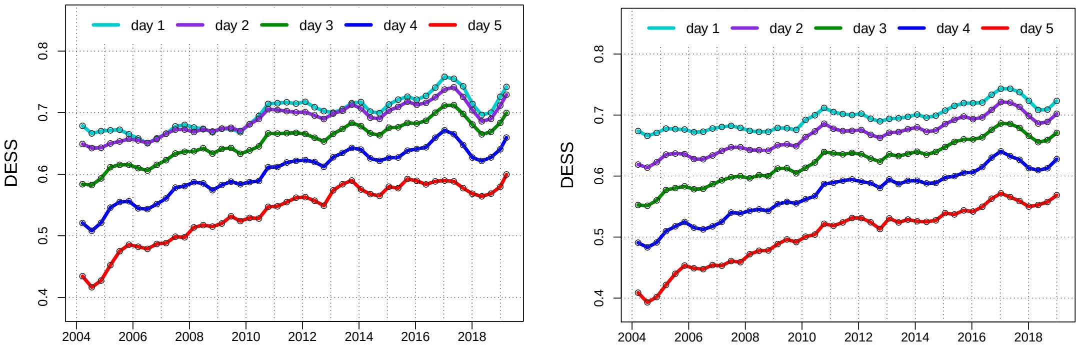

Figure 22: Long-term trends in daily precipitation ensemble forecast performance as measured by the diagonal elementary score (DESS) with (right) and without (left) accounting for representativeness uncertainty. 1-year running mean over the past 16 years. The event under focus is defined as the 95% percentile of the eigenclimatology.

allows to include the uncertainty associated with station measurements in the verification process. Finally, verification results derived with and without the perturbed ensemble approach are compared and analysed. It is shown, in summary, that accounting for observation representativeness error has a large impact on the assessment of forecast reliability, of forecast skill at short lead times, and potentially on forecast discrimination ability for rare events. These conclusions are valid for all 3 surface weather variables analysed here.

An important side benefit of this study is that it provides the basis for model-independent post-processing methods. Ensemble members (or deterministic forecasts) can be dressed with a parametric distribution using for each variable the corresponding model proposed here. Because the model fitting is based on independent observations only, it can be applied to forecasts from any model, simply adapting the parameters as a function of the model grid-spacing. For 10 m wind speed and daily precipitation, the derived probabilistic forecasts could be interpreted as valid at any given location of a model grid box. For 2 m temperature, the derived forecasts are valid at the station location because the model is based on the elevation difference between model representation and station. Importantly, the proposed approach is fully parametric, and as such straightforward to apply (by any direct model-output user or forecast provider) in order to generate as many ”members” as desired. This type of post-processing can be seen as a way to account for model limitations, due to sub-grid scale uncertainty, but cannot correct for model-specific deficiencies.

Because of their simplicity, the models proposed here could be considered as a benchmark for more complex approaches. Parameters of the generalized uncertainty models are estimated based on a European dataset, but are intended to be applied globally. Representativeness error is described in generalized terms, but each model could be developed further by considering that sub-grid scale uncertainty is weather-situation dependent or at least considering parameters that vary with seasons. More complex models would benefit both verification and potential post-processing applications.

5 Future prospects

Accounting for observation uncertainty in the verification of ensemble forecasts is planned to be performed routinely at ECMWF. Routine verification activities include monitoring the operational ensemble system performance as well as monitoring long-term trends in the form of head-line scores for example.

Routine verification also encompasses the assessment of ensemble experiments and in particular of new IFS cycle candidates. Because of its potential strong impact on the ”colour” of a scorecard (as discussed in Section 1), it is important to account for observation uncertainty in assessing cycle upgrades. This is now possible with the in-house verification software *Quaver*, which, from version 1.4.0 on, is ready to handle observation uncertainty using the perturbed ensemble approach.

Besides the variables investigated here (2m temperature, 10 m wind speed, and daily precipitation), uncertainty models for other surface and upper-air variables are also implemented in *Quaver*. For upper-air variables, random errors are considered as the main contributors to observation uncertainty and the perturbed ensemble approach is applied using normal distributions when verifying geopotential height, temperature, wind speed, and relative humidity against radiosondes. The error distribution is assumed to be independent of the value of the variable except for relative humidity where a multiplicative model is used. For upper-air wind speed, each component is treated separately. The standard deviation of the normal distributions varies for each variable as a function of the pressure level, and is estimated by data assimilation experts (Ingleby 2018, personal communication, Ingleby, 2017). Regarding ensemble wave forecast verification, observation uncertainty for significant wave height has been estimated using triple collocation technique (Abdalla *et al.*, 2011). Based on this study, an uncertainty model for this parameter will soon be implemented in *Quaver*. In contrast, uncertainty associated with cloud cover observations is difficult to assess because of the mixture of observation types and the large difference between automated and manual observation error characteristics (Mittermaier, 2012). So further investigations are required for this variable. Finally, we can note that when verification is performed against analyses, the correlation between analysis error and forecast error should be considered. The estimation of this quantity is generally not trivial (Simmons and Hollingsworth, 2002; Peña and Toth, 2014) and the use of a simple perturbed approach might not be appropriate in that case. Accounting for analysis uncertainty in forecast verification is an active field of research.

Acknowledgements

The author is grateful to Thomas Haiden, David Richardson, and Florian Pappenberger for their help in preparing this tech memo, to Nick Weber and Tom Hamill for their collaboration on a companion peer-reviewed paper (Ben Bouallègue *et al.*, 2020), to Martin Leutbecher, Philippe Lopez, and Martin Janousek for useful discussions and suggestions. A special thank goes to the author’s family for their support during the ”great lockdown”.

References

Abdalla, S., Janssen, P. A. E. M. and Bidlot, J.-R. (2011). Altimeter near real time wind and wave products: Random error estimation. *Marine Geodesy*, **34**(3-4), 393–406.

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

Ben Bouallègue, Z., Haiden, T. and Richardson, D. S. (2018). The diagonal score: definition, properties, and interpretations. *Quart. J. Roy. Meteor. Soc.*, **144**(714), 1463–1473.

Ben Bouallègue, Z., Haiden, T., Weber, N. J., Hamill, T. M. and Richardson, D. S. (2020). Accounting for representativeness in the verification of ensemble precipitation forecasts. *Monthly Weather Review*, **148**(5), 2049–2062.

Ben Bouallègue, Z., Magnusson, L., Haiden, T. and Richardson, D. S. (2019). Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quart. J. Roy. Meteor. Soc.*, **145**(721), 1741–1755.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**(1), 1–3.

Candille, G. and Talagrand, O. (2008). Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**(633), 959–971.

Duc, L. and Saito, K. (2018). Verification in the presence of observation errors: Bayesian point of view. *Quart. J. Roy. Meteor. Soc.*, **144**(713), 1063–1090.

Feldmann, K., Richardson, D. and Gneiting, T. (2019). Grid- vs. station-based postprocessing of ensemble temperature forecasts. *Geophys. Res. Lett.*, **46**.

Ferro, C. (2017). Measuring forecast performance in the presence of observation error. *Quart. J. Roy. Meteor. Soc.*, **143**(708), 2665–2676.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gneiting, T., Stanberry, L., Grimit, E., Held, L. and Johnson, N. (2008). Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test*, **17**, 211–235.

Göber, M., Zsoter, E. and Richardson, D. (2008). Could a perfect model ever satisfy a naÃ¯ve forecaster? On grid box mean versus point verification. *Met. Apps*, **15**(3), 359–365.

Haiden, T., Dahoui, M. ., Ingleby, B., de Rosnay, P., Prates, C., Kuscu, E., Hewson, T., Isaksen, L., Richardson, D., Zuo, H. and Jones, L. (2018). Use of in situ surface observations at ecmwf. *ECMWF Technical Memorandum*, **834**.

Hamill, T. and Colucci, S. (1997). Verification of eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**(6), 1312–1327.

Hamill, T. M., Hagedorn, R. and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**(7), 2620–2632.

Ingleby, B. (2017). An assessment of different radiosonde types 2015/2016. *ECMWF Technical Memorandum*, **807**.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A. and Weston, P. (2018). On the representation error in data assimilation. *Quart. J. Roy. Meteor. Soc.*, **144**(713), 1257–1278.

Jolliffe., I. T. (2017). Probability forecasts with observation error: what should be forecast? *Meteorological Applications*, **78**.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, **227**, 3515–3539.

Lopez, P., Ryu, G.-H., Sohn, B.-J., Davies, L., Jakob, C. and Bauer, P. (2011). Specification of rain gauge representativity error for data assimilation. *ECMWF Technical Memorandum*, **647**.

Massonnet, F., Bellprat, O., Guemas, V. and Doblas-Reyes, F. (2016). Using climate models to estimate the quality of global observational data sets. *Science*, **354**(6311), 452–455.

Mittermaier, M. (2012). A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**(668), 1794–1807.

Pappenberger, F., Ghelli, A., Buizza, R. and Bòdis, K. (2009). The skill of probabilistic precipitation forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications. *Journal of Hydrometeorology*, **10**(3), 807–819.

Peña, M. and Toth, Z. (2014). Estimation of analysis and forecast error variances. *Tellus A: Dynamic Meteorology and Oceanography*, **66**(1), 21767.

Pinson, P. and Hagedorn, R. (2012). Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorol. Appl.*, **19**(4), 484–500.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**(5), 1155–1174.

Röpnack, A., Hense, A., Gebhardt, C. and Majewski, D. (2013). Bayesian model verification of nwp ensemble forecasts. *Mon. Weather Rev.*, **141**(1), 375–387.

Saetra, O., Hersbach, H., Bidlot, J.-R. and Richardson, D. (2004). Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Weather Rev.*, **132**(6), 1487–1501.

Santos, C. and Ghelli, A. (2012). Observational probability method to assess ensemble precipitation forecasts. *Q. J. R. Meteorolog. Soc.*, **138**(662), 209–221.

Scheuerer, M. and Hamill, T. (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.*, **143**(11), 4578–4596.

Simmons, A. J. and Hollingsworth, A. (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **128**(580), 647–677.

Thorarinsdottir, T. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A*, **173**, 371–388.

Wilks, D. and Vannitsem, S. (2018). Uncertain forecasts from deterministic dynamics. In S. Vannitsem, D. Wilks and J. Messner (Eds), *Statistical Postprocessing of Ensemble Forecasts 1st Edition*, pp. 1–13, Elsevier.

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences*. 2nd Edn. Academic Press, New York, 627pp.

Yamaguchi, M., Lang, S. T. K., Leutbecher, M., Rodwell, M. J., Radnoti, G. and Bormann, N. (2016). Observation-based evaluation of ensemble reliability. *Quart. J. Roy. Meteor. Soc.*, **142**, 506–514.

APPENDIX

Complementary figures

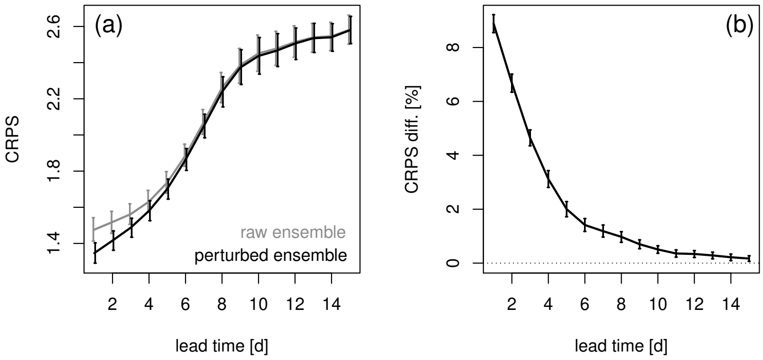

Figure 23: (a) CRPS (in K) estimated with (black) and without (grey) accounting for representativeness uncertainty and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Results valid for ***2 m temperature*** *ensemble forecasts, Europe,* ***Winter*** *2018-2019.*

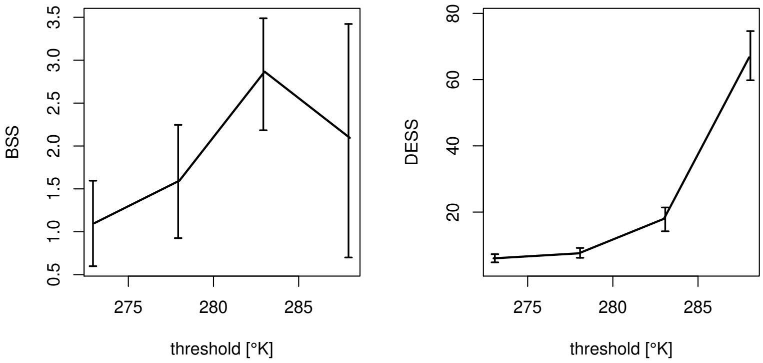

Figure 24: (a) Brier skill score (BSS), and (b) diagonal elementary skill score (DESS) as a function of the event-threshold. Results valid for ***2 m temperature*** *ensemble forecast at* ***day 5****, Europe,* ***Winter*** *2018-2019.*

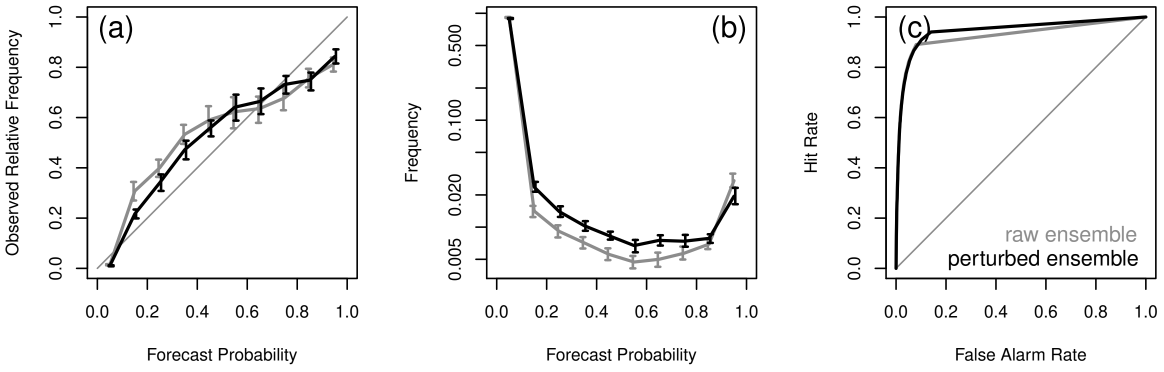

*Figure 25: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of **283** K. Results valid for **2 m temperature** ensemble forecast at **day 5**, Europe, **Winter** 2018-2019.*

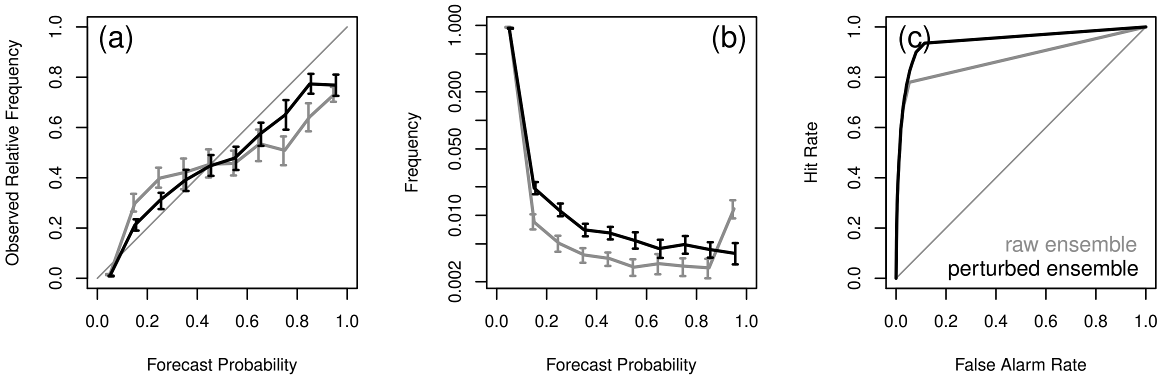

*Figure 26: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of **298** K. Results valid for **2 m temperature** ensemble forecast at **day 5**, Europe, **Summer** 2018-2019.*

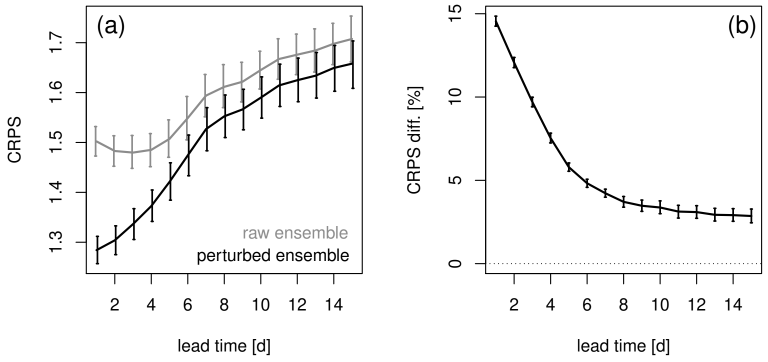

*Figure 27: (a) CRPS (in m/s) estimated with (black) and without (grey) accounting for representativeness uncertainty and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Results valid for **10 m wind speed** ensemble forecasts, Europe, **Winter** 2018-2019.*

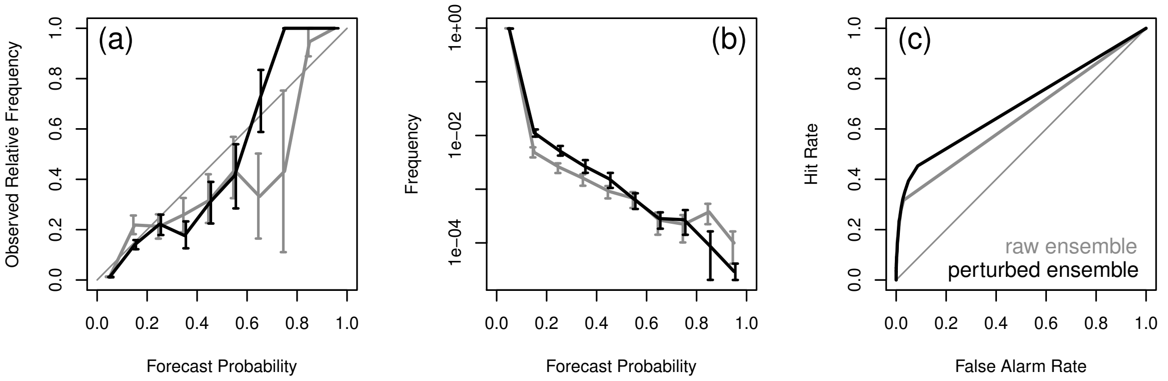

*Figure 28: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of **10** m/s. Results valid for **10 m wind speed** ensemble forecast at **day 5**, Europe, **Summer** 2018.*

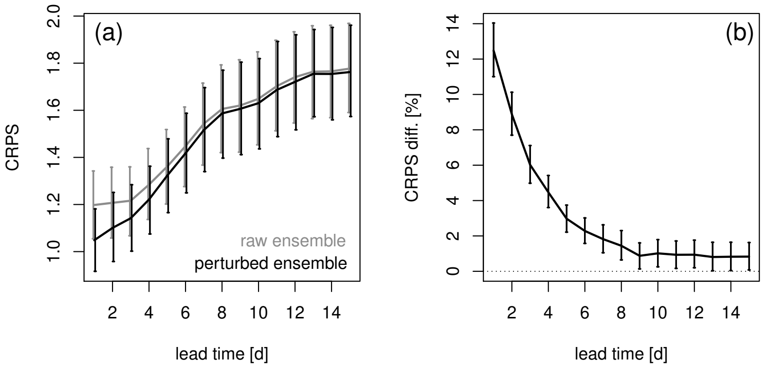

*Figure 29: (a) CRPS (in mm/24h) estimated with (black) and without (grey) accounting for representativeness uncertainty and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Results valid for **daily precipitation** ensemble forecasts, Europe, **Winter** 2018-2019.*

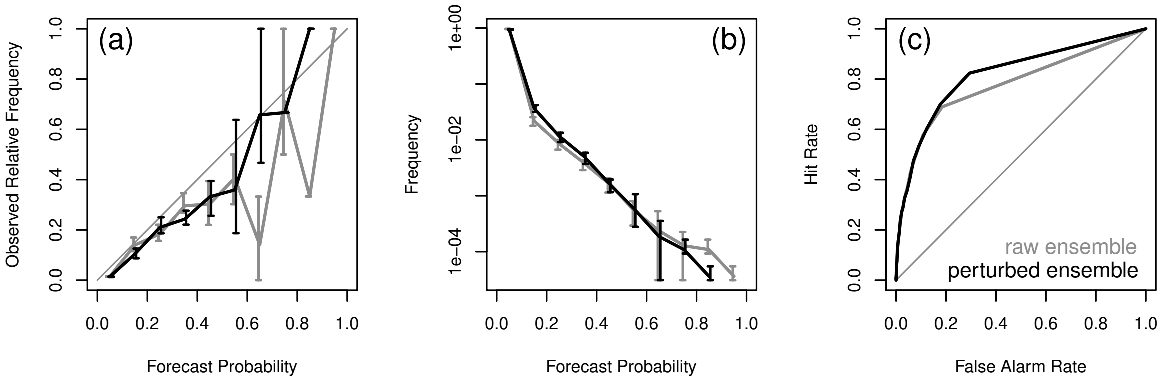

*Figure 30: (a) Reliability diagram, (b) sharpness diagram, (c) and ROC curve for an event-threshold of **20** mm/24h. Results valid for **daily precipitation** ensemble forecast at **day 5**, Europe, **Summer** 2018. (Published as Fig. 10 in Ben Bouallègue et al. (2020). © American Meteorological Society. Used with permission.)*