



# Novel multivariate quantile mapping methods for ensemble post-processing of medium-range forecasts

Kirien Whan<sup>a,\*</sup>, Jakob Zscheischler<sup>b,c,d</sup>, Alexander I. Jordan<sup>e,f</sup>, Johanna F. Ziegel<sup>c,f</sup>

<sup>a</sup> The Royal Netherlands Meteorological Institute, De Bilt, the Netherlands

<sup>b</sup> Climate and Environmental Physics, University of Bern, Switzerland

<sup>c</sup> Oeschger Centre for Climate Change Research, University of Bern, Switzerland

<sup>d</sup> Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

<sup>e</sup> Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

<sup>f</sup> Institute of Mathematical Statistics and Actuarial Sciences, University of Bern, Switzerland

## ARTICLE INFO

### Keywords:

Statistical post-processing  
Ensemble model output statistics  
Compound events  
Quantile mapping

## ABSTRACT

Statistical post-processing is an indispensable tool for providing accurate weather forecasts and early warnings for weather extremes. Most statistical post-processing is univariate, with dependencies introduced via use of an empirical copula. Standard empirical methods take a dependence template from either the raw ensemble output (ensemble copula coupling, ECC) or the observations (Schaake Shuffle, SSH). There are drawbacks to both methods. In ECC it is assumed that the raw ensemble simulates the dependence well, which is not always the case (e.g. 2-meter temperature in The Netherlands). The Schaake Shuffle is not able to capture flow dependent changes to the dependence and the choice of observations is key. Here we compare a reshuffled standard ensemble model output statistics (EMOS) approach with two multivariate bias adjustment approaches that have not been used before in a post-processing context: 1) the multivariate bias correction with  $N$ -dimensional probability density function transform (MBCn) and 2) multivariate ranks that are defined with optimal assignment (OA). These methods have the advantage that they are able to explicitly capture the dependence structure that is present in the observations. We apply ECC, the Schaake Shuffle, MBCn and OA to 2-m and dew point temperature forecasts at seven stations in The Netherlands. Forecasts are verified with both univariate and multivariate methods, and using a heat index derived from both variables, the wet-bulb globe temperature (WBGT). Our results demonstrate that the spatial and inter-variable dependence structure is more realistic in MBCn and OA compared to ECC or the Schaake Shuffle. The variogram score shows that while ECC is most skilful for the first two days, at moderate lead times MBCn is most skilful and at the longest lead times OA is more skilful than both ECC and MBCn. Overall, we highlight the importance of considering the dependence between variables and locations in the statistical post-processing of weather forecasts.

## 1. Introduction

Many aspects of society require skilful weather forecasts to make planning decisions. It is essential that these forecasts are issued with sufficient lead time so that the required actions can be taken. Weather forecasts contain unavoidable errors, that increase with forecast lead time, due to errors in the initial conditions and the parameterisation of sub-grid scale processes (Palmer et al., 2005). Ensemble forecasts communicate the forecast uncertainty to users and allow them to make better decisions, compared to users with only a deterministic forecast (Joslyn and LeClerc, 2012).

Statistical post-processing (or model output statistics) uses the relationship between a historical set of forecasts and the corresponding observation to correct systematic biases in new forecasts. The current benchmark method for the calibration of ensemble forecasts is ensemble model output statistics (EMOS), a univariate post-processing method that was first developed for temperature (Gneiting et al., 2005), and has been applied to many other variables including precipitation (e.g. Scheuerer and Hamill, 2015a; Whan and Schmeits, 2018; van Straaten et al., 2018) and wind speed (e.g. Lerch and Thorarinsdottir, 2013). EMOS returns the parameters of the forecast distribution for each new forecast, for example the mean and standard deviation of a normal

\* Corresponding author.

E-mail address: [whan@knmi.nl](mailto:whan@knmi.nl) (K. Whan).

<https://doi.org/10.1016/j.wace.2021.100310>

Received 29 May 2020; Received in revised form 29 January 2021; Accepted 4 February 2021

Available online 12 February 2021

2212-0947/© 2021 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

distribution for a temperature forecast. For many applications individual ensemble members are required, rather than the parameters of the forecast distribution, for example, as the meteorological inputs for a hydrological model. We can draw ensemble members from the forecast distribution, as either equally spaced or random quantiles. However, the production of ensemble members in this way destroys any inter-variable, spatial or temporal correlations that are present in reality and the raw ensemble. The benchmark methods for the reintroduction of dependence structures are ensemble copula coupling (ECC, Schefzik et al. (2013)), where the dependence template is taken directly from the raw ensemble, or the Schaake Shuffle (SSh, Clark et al. (2004)), where the template is taken from observations. There are advantages and disadvantages of each method. ECC is able to capture flow dependencies in the correlations, but its ability to accurately represent dependence structures depends solely on the ability of the raw ensemble to simulate these dependencies, which is not always the case to a sufficient degree. For the SSh, the selection of a set of days from the observations is crucial (Wilks, 2015). Generally, flow dependent changes in the correlation structure cannot be captured, although perhaps this is possible with a very specific selection of days from the observational archive (Scheuerer et al., 2017). It is clear that there are drawbacks to both methods. Other parametric methods to model the bivariate dependence have been suggested (e.g. Pinson (2012)) but they are often not suited to non-Gaussian cases, or they are not feasible in high dimensions (Wilks, 2015), an extension to more dimensions was introduced by Keune et al. (2014).

Most post-processing approaches focus on a single variable. When issuing weather warnings, this means that only single hazards can be predicted with high confidence. However, it is becoming increasingly clear that multivariate hazards have particularly large impacts on society (Leonard et al., 2014; Zscheischler et al., 2018). Impact-relevant dependencies may occur between meteorological variables across space or time (Zscheischler et al., 2020), which needs to be considered in the post-processing of multivariate weather hazard forecasts. Purely univariate adjustments of driving variables may even increase biases in multivariate hazard indicators (Zscheischler et al., 2019). This means that we need methods that are better able to capture or adjust for the relevant dependence structures.

Quantile mapping is a general bias-correction method that maps the quantiles from a forecast to those of the observations (Hopson and Webster, 2010; Maraun, 2013; Voisin et al., 2010). It is a simple and popular method for the bias-correction of climate model output, but is used little in statistical post-processing of numerical weather forecasts in favour of more advanced methods like EMOS. Some recent developments in quantile mapping methods, developed for correcting biases in climate model simulations, have demonstrated the ability to better transfer the dependence structures of the observations to the target data set (Cannon, 2018; Robin et al., 2019; François et al., 2020). Multivariate bias-correction with  $N$ -dimensional probability density function transform (MBCn) from Cannon (2018), and optimal assignment (OA), in the spirit of Chernozhukov et al. (2017) and similar to Robin et al. (2019), are described in more detail below (Section 2.2).

MBCn is a method from computer vision, that Cannon (2018) demonstrated in two regional climate model (RCM) bias-correction applications. First, Cannon (2018) bias-corrected seven variables, including precipitation, wind speed, surface temperature and relative humidity, in order to calculate the Canadian Forest Fire Weather Index (FWI). The bias-correction was either carried out individually for each marginal distribution using univariate quantile mapping, or for all variables together using MBCn. Application of these methods for precipitation required consideration of precipitation's probability mass at zero and a transformation was required for bounded variables. Both univariate quantile mapping and MBCn improved upon the simulation of the FWI compared to the raw RCM output, but errors were substantially smaller when the inter-variable dependencies were taken into account with MBCn. Second, Cannon (2018) applied quantile mapping to the precipitation field of a RCM to assess the ability of the methods to

**Table 1**

The seven stations in The Netherlands used in this study.

Station Name	Station Number	Latitude (° N)	Longitude (° E)
De Kooy	235	52.93	4.78
Schiphol	240	52.32	4.78
De Bilt	260	52.10	5.18
Eelde	280	53.12	6.58
Twenthe	290	52.27	6.88
Vlissingen	310	51.45	3.60
Maastricht	380	50.90	5.77

capture spatial dependencies. It is shown that MBCn is best able to match the observed precipitation amounts and the spatial correlation structure. The raw RCM has a spatial correlation structure that is too strong, and this is retained by univariate quantile mapping. After applying MBCn, the spatial dependence is realistically simulated. Robin et al. (2019) demonstrate the use of optimal transport theory to construct a joint distribution that can be used to extend the univariate quantile mapping to the multivariate case. They use a simulated example to show that inter-variable correlation structure is best corrected by their optimal transport method, before applying the method to temperature and precipitation at 12 locations in France (Robin et al., 2019).

Here, we compare the benchmark calibration method, EMOS, with the use of an empirical copula (either ECC or SSh), to the two multivariate bias-correction methods that have been designed to correct dependence structures across different dimensions (MBCn and OA). As an example application we use the spatial and inter-variables dependence of 2-m temperature and dew point temperature at seven stations in The Netherlands.

## 2. Data and methods

### 2.1. Data

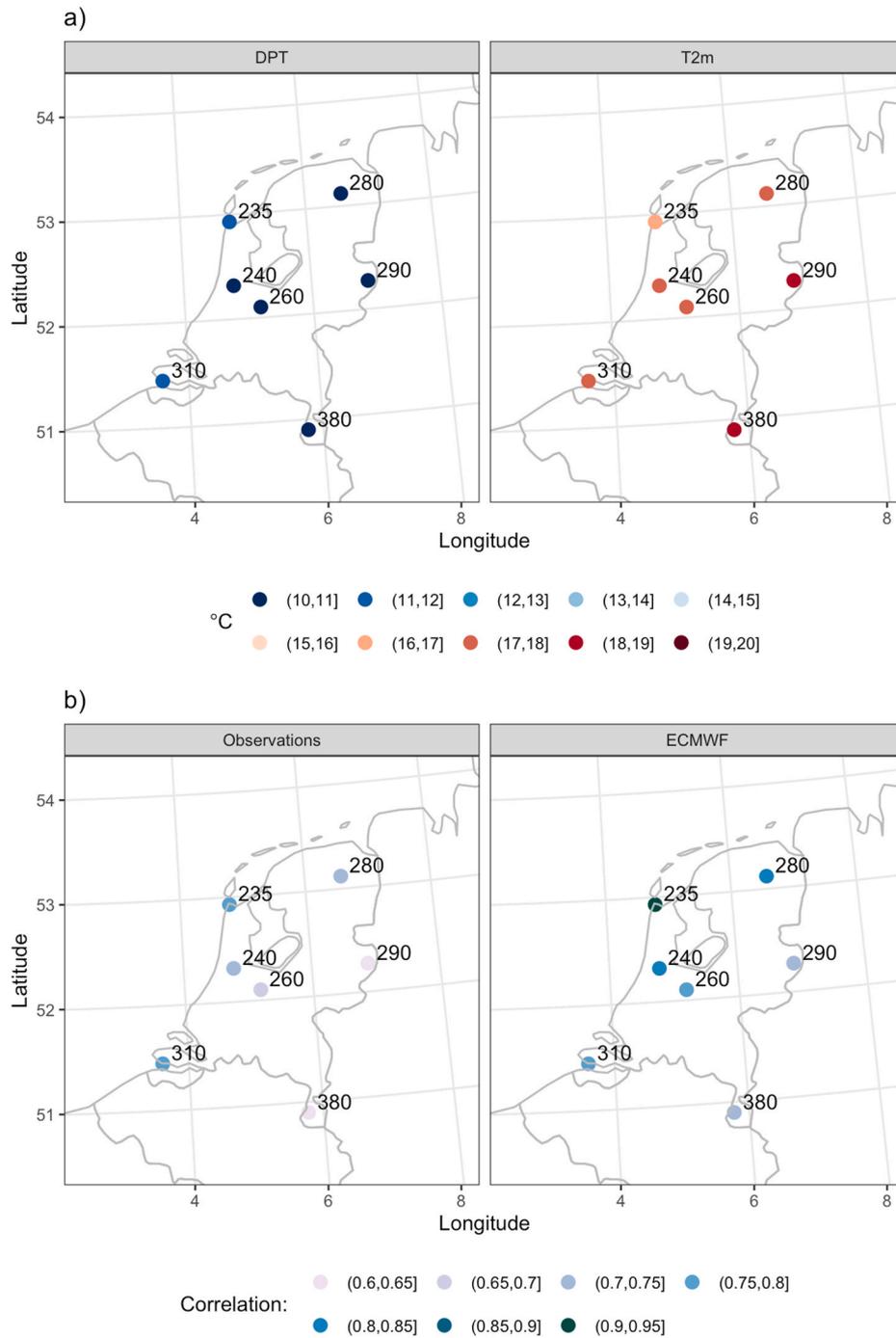
The target observations are extended summer (mid-April to mid-October) 2m temperature (T2m) and dew point (DPT) station observations at seven stations in the Netherlands (Table 1 and Fig. 1), over the period 2011–2018. In the Netherlands, DPT are, on average, higher on the coast and lower inland (Fig. 1a). T2m displays the opposite pattern, with higher average values in the west and south (particularly at stations 290 and 380) and the lowest average values in De Kooy (station 235), which is located near the ocean (Fig. 1a).

The 51 members of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system are initialised each day at 00 UTC, and are valid at 12 UTC on each day from day 1 (+12 h) to day 10 (+228 h). There have been several model changes during the period 2011–2018. A recent study has shown that the bias in the post-processed forecasts introduced by using several model versions is compensated for by the reduction in variance in the post-processed forecasts that is achieved by using more data (Lang et al., 2019). While the variance in the fitted model parameters is reduced with larger training sample sizes, additional biases may be introduced as the error characteristics of the NWP can change with model updates. The raw forecasts are the closest grid-cell to the station location.

We compare the skill of the post-processing methods using cross-validation. We use four-fold cross-validation, so that we train on 6 years and test on the remaining 2 years. We post-process all forecasts to include spatial and inter-variable dependencies. We then verify all years together using the univariate and multivariate verification measures outlined in Section 2.4.

### 2.2. Post-processing methods

We compare five post-processing methods: Ensemble model output statistics (EMOS) (Gneiting et al., 2005), EMOS forecasts that have the dependencies restored with the Schaake shuffle (SSh) (Clark et al., 2004)



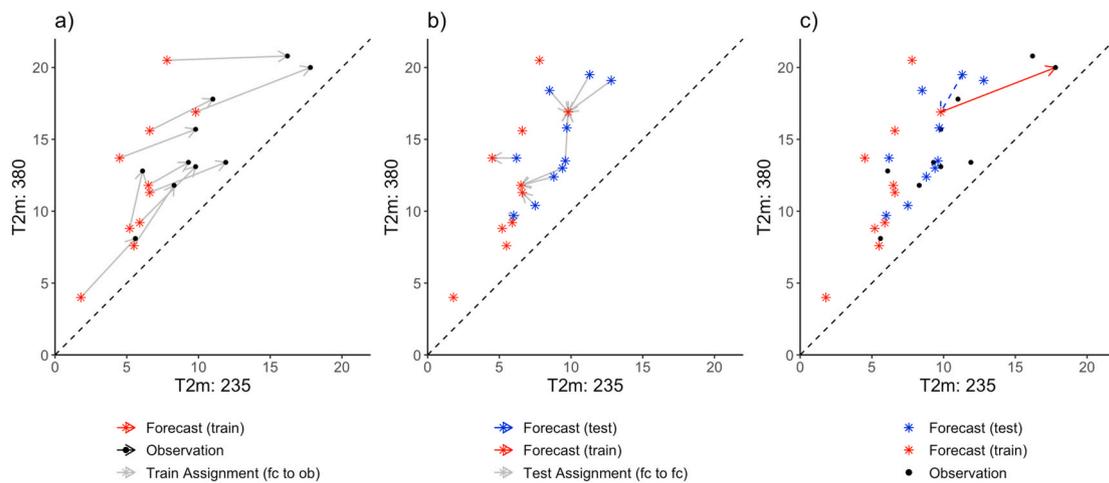
**Fig. 1.** a) The climatology of dew point temperature (DPT, left) and 2m temperature (T2m, right) at seven stations in the Netherlands. b) The correlation between T2m and DPT in observations (left) and at the +12 h forecast lead time in all members of the ECMWF ensemble (right). The correlations are similar for other forecast lead times. Numbers indicate the station identifiers that can be seen in Table 1.

or with ensemble copula coupling (ECC) (Scheffzik et al., 2013), multivariate bias-correction (MBCn) (Cannon, 2018), and optimal assignment (OA) inspired by Chernozhukov et al. (2017); Robin et al. (2019). A short description of each is given below.

### 2.2.1. Ensemble model output statistics (EMOS) with an empirical copula (SSH or ECC)

The benchmark method is a univariate EMOS in combination with ECC and SSH, as baseline methods. We assume that temperature and dew point temperature follow a normal distribution. We use the ‘gamlss’ package in R (Rigby and Stasinopoulos, 2005).

We let the parameters of the distribution depend linearly on the ensemble, so that, for example, the mean is a function of the ensemble mean. We fit a local model for the extended summer period to each station and variable separately, and take the quantiles at probabilities  $i/52, i = 1, \dots, 51$ , from the forecast distribution to match the number of members in the raw ensemble. These forecasts are then reordered to restore dependencies. This ordering is taken from a ‘dependence template’ that can come from a number of sources, as described below. Our default choice for multivariate EMOS is a dependence template assuming perfect positive correlation to serve as a baseline. We call this ‘EMOS-ppc’. We use ECC and SSH to estimate the dependencies from the



**Fig. 2.** An example of the optimal assignment algorithm. a) First, forecasts (red stars) are mapped to their optimally assigned observation (black circles) during the training period. b) Next, forecasts during the test period (blue stars) are mapped to their closest forecast in the training period (red stars). c) The path for the correction of a single forecast, from test forecast to train forecast (blue dashed line) and from train forecast to observation (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

data. ECC takes the dependence template directly from the raw ensemble (Schefzik et al., 2013). For the SSH we take a similar approach to Clark et al. (2004). We draw 51 random days that occur during the study period, within 5 days (before or after) the target day, and in a different year than the target day. We use functions from the ‘depPPR’ package in R to restore dependencies with ECC and SSH (Saunders and Whan, 2020).

### 2.2.2. Multivariate bias-correction (MBCn)

MBCn is a multivariate extension of quantile mapping that was developed for computer vision (Pitié et al., 2005; Pitié et al., 2007) and extended for use by the climate community by Cannon (2018). Quantile mapping makes no assumptions about the marginal distribution. An example and description of the MBCn algorithm is given in Section 3 in Cannon (2018) (his Figs. 1 and 2). This example demonstrates the benefit of quantile mapping that takes the dependencies between variables into account. MBCn is an iterative method that first applies a random orthogonal rotation, that could be constructed via QR decomposition of normally distributed random values, to the training and test set, then adjusts the marginal distributions with quantile mapping and then inverts the rotation. This continues until the multivariate distribution of the training set matches the target distribution of the observations, or the maximum number of iterations is reached (Cannon, 2018). We use the default maximum number of iterations of 30. We apply MBCn in a 14 dimensional setting to all stations (seven) and variables (two) at once.

We use the ‘MBC’ package in R for MBCn (Cannon, 2018). We compared adjusting each ensemble member individually and all members together. Quantile mapping approaches require three data sets, 1. the observations in the training period, 2. forecasts in the training period, and 3. forecasts in the test period. We looped over ensemble members, so that each member was used for the training and test sets. However, as ECMWF ensemble members are interchangeable it makes little difference which ensemble member is used in the training set and which is used in the test set. Indeed, the results are similar if we use a random member for the training set and loop over all members in the test set.

When adjusting all members together the number of observations remains the same (i.e. the number of cases in the training set) but the number of forecasts increases, as we use all members together. The number of forecasts in the training set is the number of training days times the number of ensemble members, and the number of forecasts in the test set is the number of days in the test set times the number of

ensemble members.

Given the large training data sets, there are no large differences between the approaches, as the bias is equally well-estimated from the smaller data set provided by each member and the combined data set containing all members.

### 2.2.3. Optimal assignment (OA)

OA is another multivariate generalization of quantile mapping, based on the interpretation of quantile mapping as an optimal transport problem. Like MBCn, there are no distributional assumptions made in OA. In the empirical context of a finite training data set, optimal transport is equivalent to optimal assignment. Each forecast in the training set is linked to an observation in the training set by a one-to-one relationship (Fig. 2a). These links are found for each ensemble member individually, and the assignment is optimal in the sense of minimizing the total sum of squared Euclidean distances in assigned forecast-observation pairs. As in the MBCn approach, any forecast or observation is a 14-dimensional vector comprising values of two variables for seven stations, therefore the Euclidean distances are calculated in 14-dimensional space. In a nutshell, OA minimizes the average distance that the forecasts in the training set need to be moved such that the set of observations is recovered.

The post-processing of a single forecast in the test data set maps such a forecast to its analogue in the training set by finding the closest forecast in terms of Euclidean distance. Multiple forecasts in the test period can be mapped to a single forecast in the training period, as shown in Fig. 2b. Finally, the optimal assignment map estimated from the training set is applied (Fig. 2c). In principle, OA is not limited in its dimensions or the marginal distributions, although in practice computational constraints and the appropriateness of the distance measure must be considered. Since temperature is regular in that it is well-described by a Gaussian distribution, the Euclidean distance or weighted versions should be appropriate, which may not be the case for precipitation or wind speed.

The optimal assignment is found using, as input, all pairwise distances between forecasts and observations in the training set. Since Euclidean distance computations scale linearly in the number of dimensions, the dimensionality when considering multiple stations or lead times is only indirectly limited by the requirement of a larger training set to accurately represent the underlying process. For OA, the size of the training set does impose computational constraints. Therefore, the optimal assignment is estimated on a member-by-member basis. We use the ‘clue’ package in R (Hornik, 2005, 2019) for its implementation of

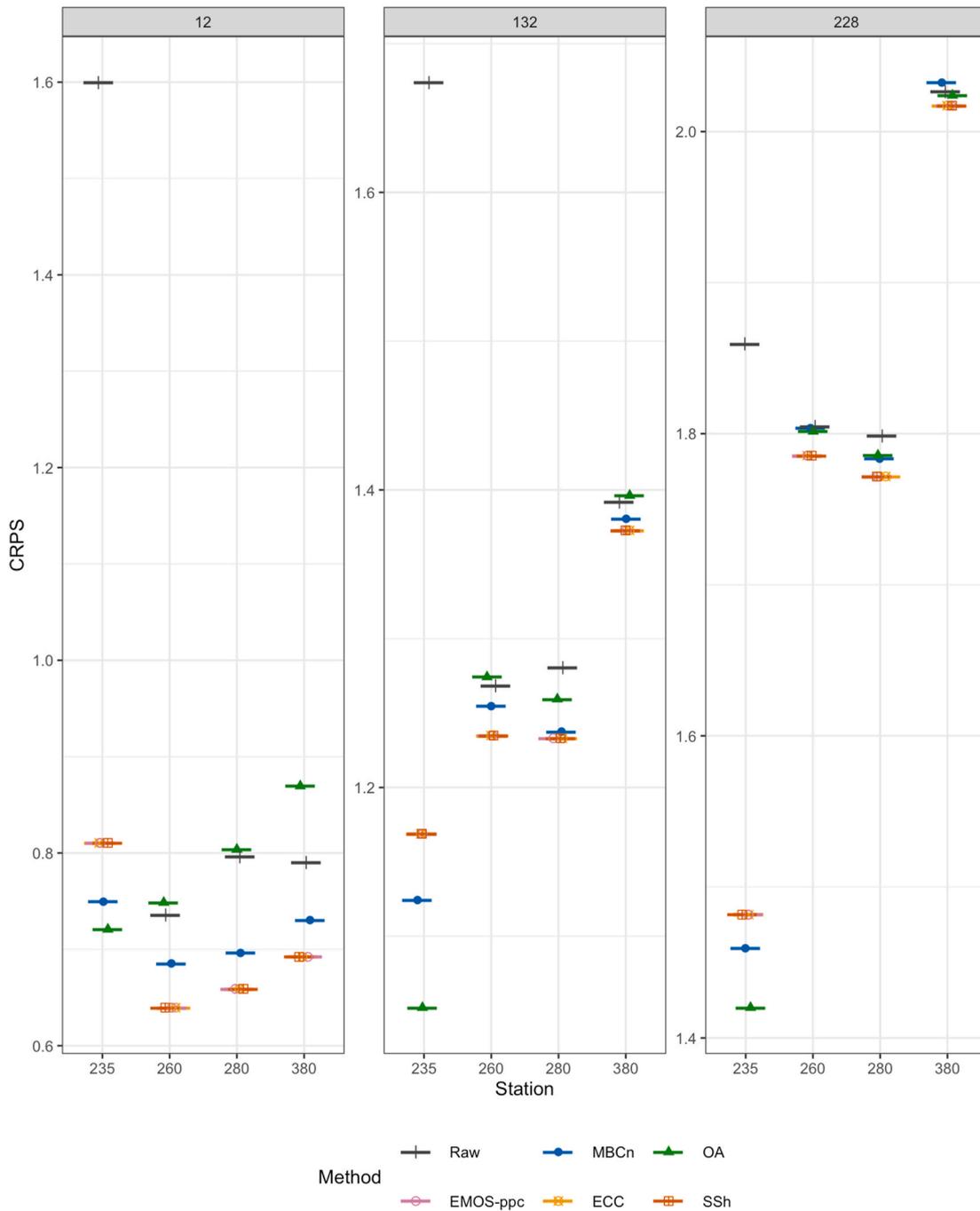


Fig. 3. The mean CRPS of T2m forecasts at De Kooy (235), De Bilt (260), Eelde (280), and Maastricht (380) at the +12, +132 and + 228 h lead times.

the Hungarian algorithm for optimal assignment, which has a cubic computational complexity and quadratic memory requirement in the number of observations. As such, estimation including all 51 ensemble members increases runtime by a factor of over one hundred thousand and memory usage by a factor of over two thousand compared to single-member estimation. The reduction to an arbitrary single member is unproblematic, since the ECMWF ensemble members are exchangeable by design. It is possible to increase the ensemble size by adjusting each member by the optimal assignments of all other members. This results in an ensemble of 2 601 members ( $51 \times 51$ ). The variogram score is largely similar to that shown below, but it is slightly lower. This decrease in the variogram score is likely due to the additional ensemble members resulting a more smooth forecast distribution, rather than a real increase in skill that stems from a better estimation of the error

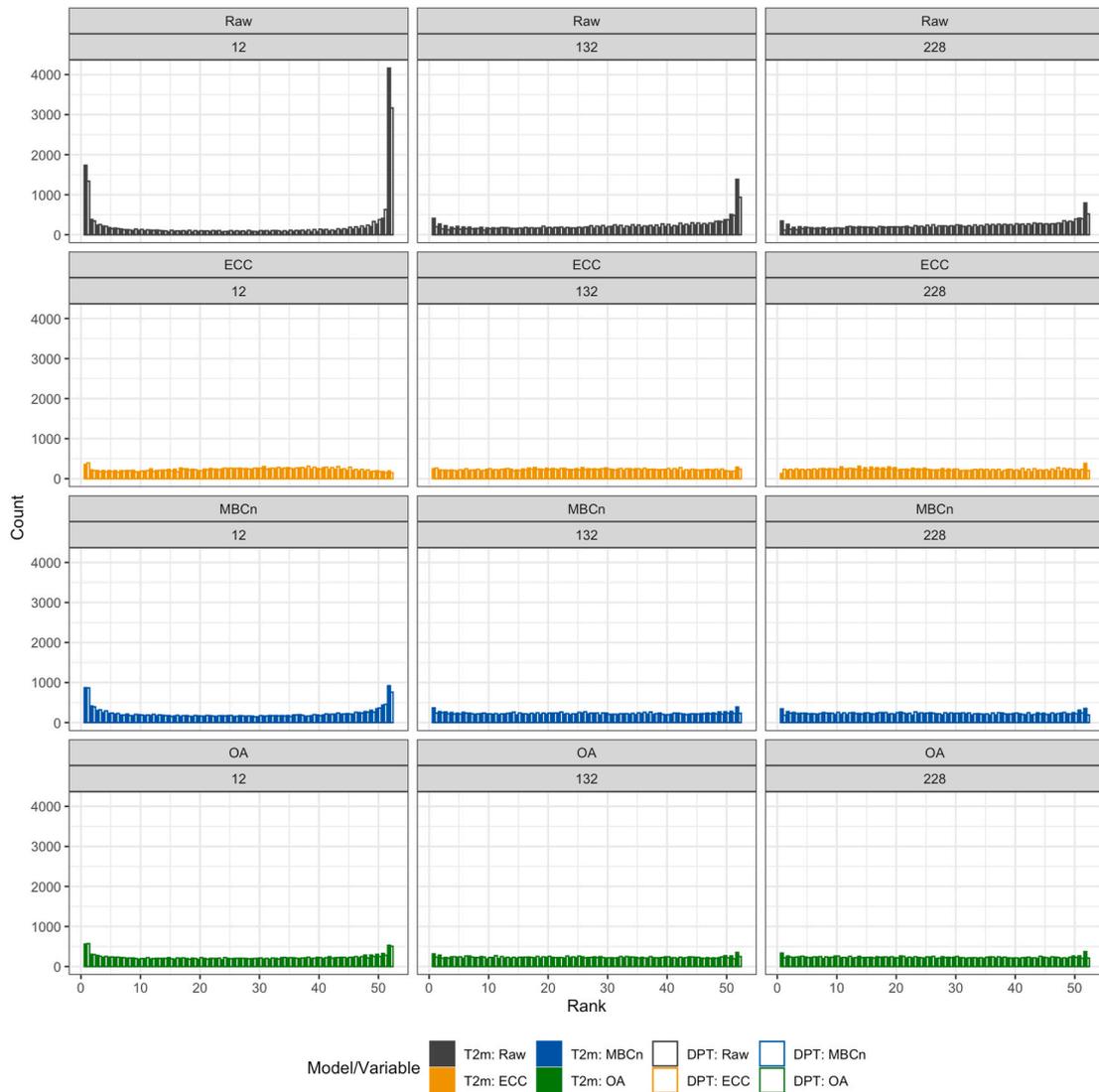
characteristics of the forecasts.

### 2.3. Heat index

Simulating the correct dependence structure is important for the calculation of indices that depend on both variables. We use the wet-bulb globe temperature (WBGT) as a heatwave index that is calculated from T2m and DPT, according to Lemke and Kjellstrom (2012). Further details can be found in Appendix B. We calculate the WBGT from T2m and DPT from the raw and post-processed ensemble forecasts.

### 2.4. Verification measures

We use standard univariate and multivariate scores to verify the raw



**Fig. 4.** Rank histograms for all stations from the Raw ECMWF forecast (black), and forecasts post-processed with ECC (yellow), MBCn (blue), and OA (green) at +12, +132 and +228 h lead times. T2m forecasts are in solid bars and DPT forecasts are in open bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and post-processed forecasts. We use the continuous ranked probability score (CRPS) for univariate verification (eq. (1)). We score the forecasts with a univariate score after we have post-processed all forecasts with the dependencies, i.e. we have restored the dependence structure with ECC or the SSh, and we have used a 14 dimensional vector for each forecast and observation in MBCn and OA. The CRPS is a proper score that measures the difference between a forecast distribution and a real-valued observation,

$$CRPS(F, y) = \mathbb{E}_F |X - y| - \frac{1}{2} \mathbb{E}_F |X - X'|, \quad (1)$$

where  $F$  is the forecast distribution,  $y$  is the observation, and both  $X$  and  $X'$  are independent random variables with distribution  $F$ .

For multivariate verification we use the energy score (eq. (2)), which is a multivariate generalisation of the CRPS,

$$ES(F, y) = \mathbb{E}_F \|X - y\| - \frac{1}{2} \mathbb{E}_F \|X - X'\|, \quad (2)$$

where  $\|\cdot\|$  denotes the Euclidean norm (Gneiting et al., 2008). However, the energy score can be insensitive to certain errors in the correlation structure, so we also use the variogram score (eq. (3)), which is more

sensitive to misspecifications in the dependence structure (Scheuerer and Hamill, 2015b),

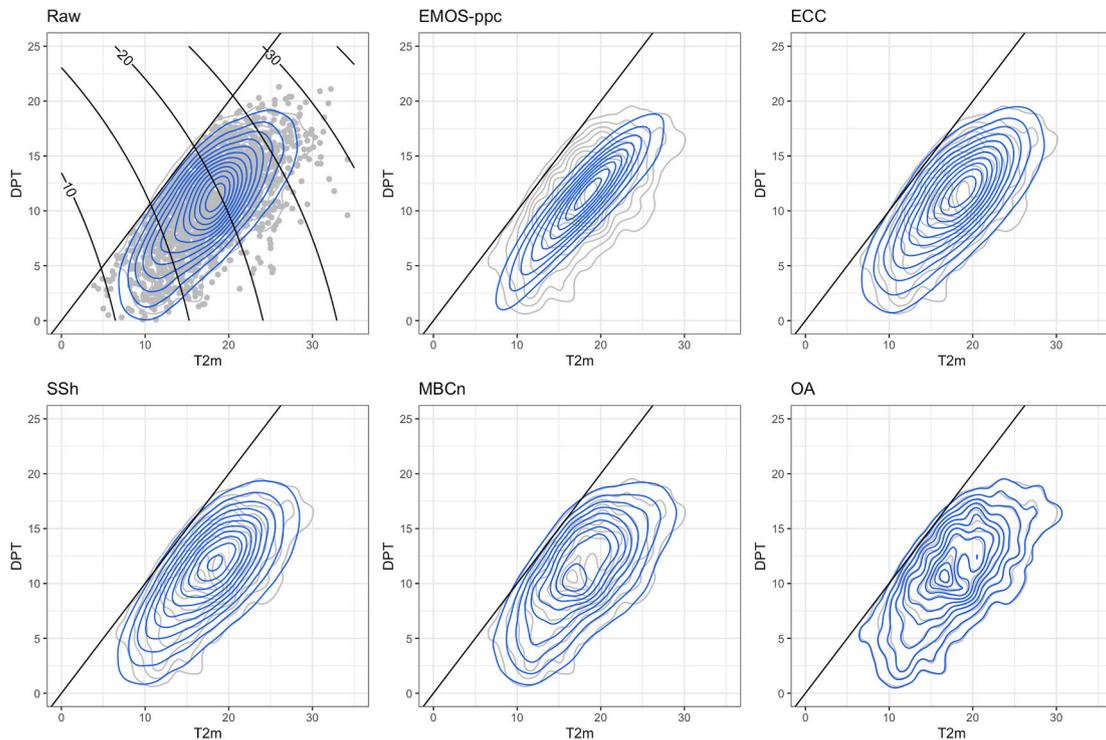
$$VS^p(F, y) = \sum_{i,j=1}^d (|y_i - y_j|^p - E_F |X_i - X_j|^p)^2, \quad (3)$$

where  $X_i$  and  $X_j$  are the  $i^{th}$  and the  $j^{th}$  component of a random vector  $X$  (Scheuerer and Hamill, 2015b). We use  $p = 0.5$ , but sensitivity testing showed that results are similar for  $p = 1$  and  $p = 2$ .

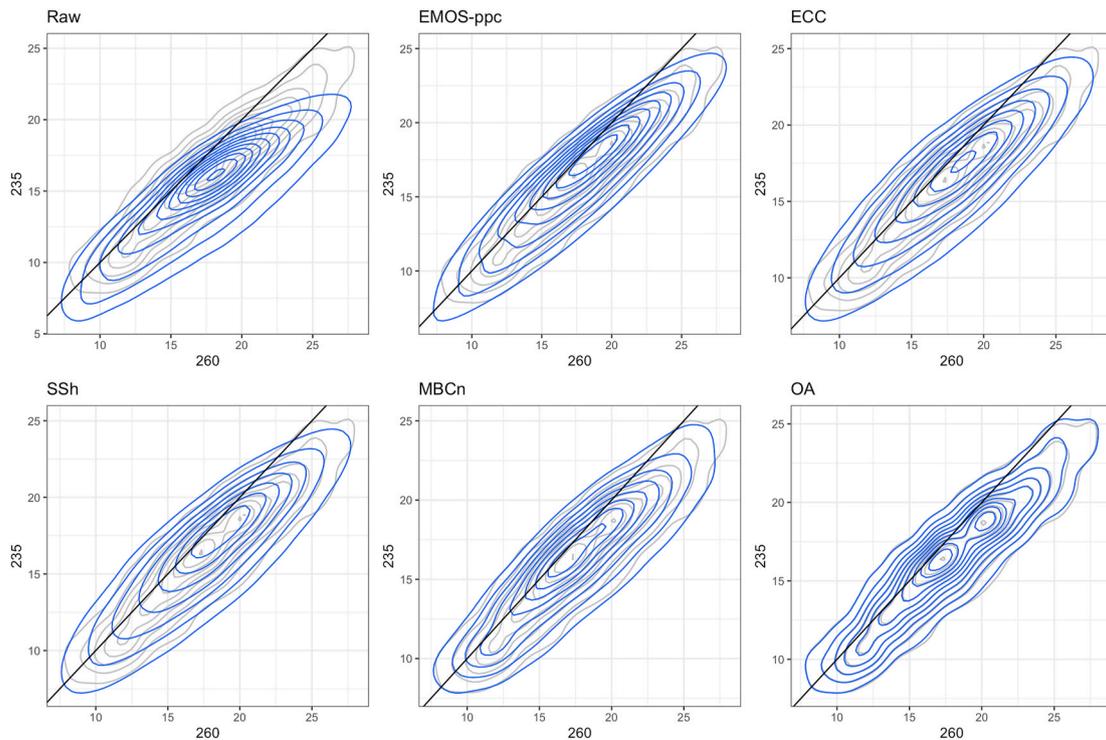
We calculate all scores using the ‘scoringRules’ package in R (Jordan et al., 2019).

### 3. Results

We first assess the ability of the raw ECMWF ensemble to simulate the inter-variable dependence at each station in The Netherlands (Fig. 1b shows the +12 h lead-time and Figure A.12 shows all lead-times.). There is a distinct longitudinal effect in observations, with stronger correlations ( $> 0.75$ ) between T2m and DPT evident on the coast (e.g. De Kooy and Vlissingen, stations 235 and 310), and weaker correlations ( $< 0.75$ ) found inland (e.g. Twenthe and Maastricht, stations 290 and 380). ECMWF over-estimates the correlation between



**Fig. 5.** A 2d density plot of T2m and DPT at De Bilt. Grey = observations, blue = the +228 h forecast. Black lines show isolines of the heat index that is calculated from T2m and DPT. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

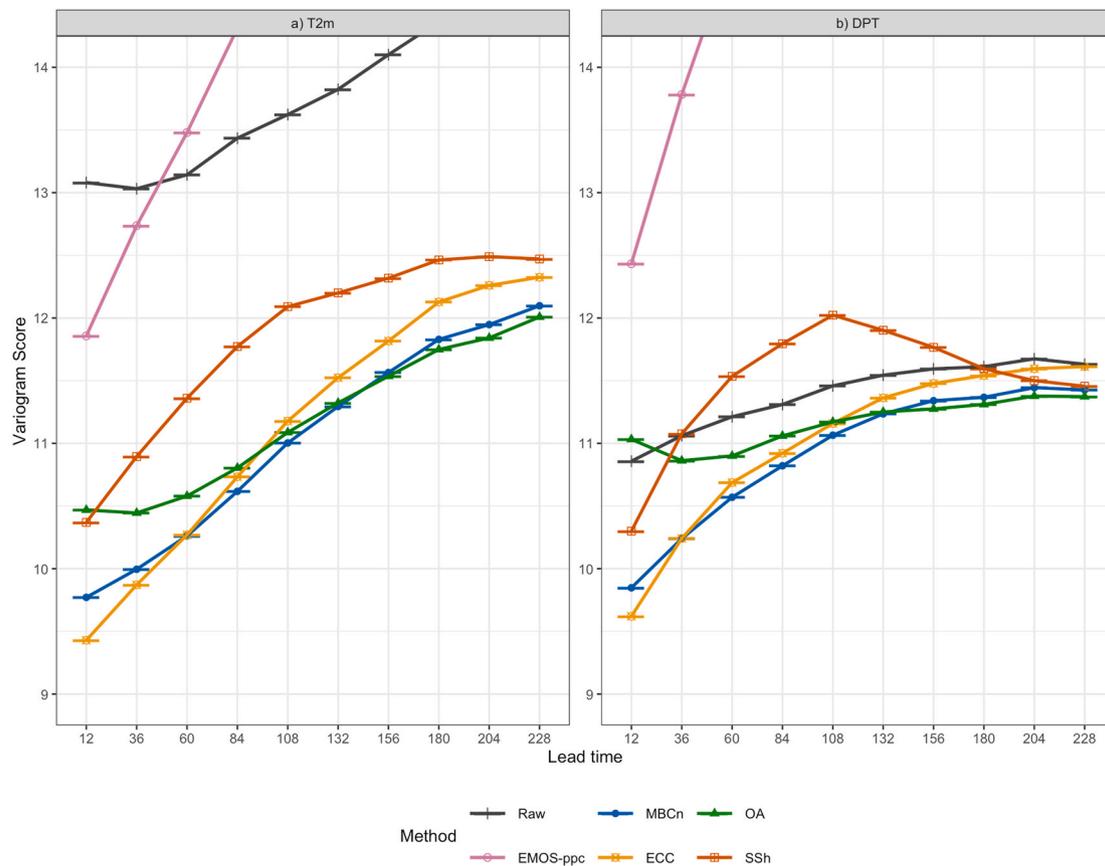


**Fig. 6.** A 2d density plot of T2m at De Bilt (station number 260) and De Kooy (station number 235). Grey = observations, blue = the +228 h forecast. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

T2m and DPT at all stations, including Vlissingen although the difference here is small, and at all lead times compared to observations. De Kooy (station number 235) is particularly interesting. In ECMWF there is a very strong correlation ( $> 0.9$ ) between T2m and DPT at De Kooy that is not seen in observations (0.75 – 0.8). This is likely due to location of

De Kooy on the coast the large maritime influence, and the fact that the closest grid cell to the station location is the ocean.

In a univariate verification setting, there are no differences between the EMOS based methods, i.e. EMOS-ppc, SSh, and ECC (Fig. 3). EMOS is the most skilful in terms of the CRPS for univariate calibration, except



**Fig. 7.** The mean variogram score ( $p = 0.5$ ) of spatial forecasts (all stations) for a) T2m and b) DPT. The variogram score for the raw ensemble and EMOS-ppc extends off the top of the figure.

for in De Kooy. In De Kooy, where the errors in the raw model are very large, EMOS is not as successful in correcting these errors and it is clear that OA and MBCn have a smaller CRPS. Calibration of the univariate forecasts is assessed with rank histograms. The raw forecast is under-dispersed at short lead times and has a cool bias. The EMOS forecasts are well calibrated at all lead times. MBCn and OA forecasts correct most of the under-dispersion at short lead times although forecasts are not perfectly calibrated, particularly for MBCn. Forecasts are well calibrated at longer lead times for both MBCn and OA (Fig. 4).

The correlation structure between variables or between locations is best adjusted by MBCn and OA (Fig. 5 and Fig. 6, respectively). The density contours of the joint distribution of the variables in De Bilt are far from elliptical shapes (wavy grey lines in Fig. 5). The estimated joint distributions of variables with EMOS-ppc and ECC are too close to an elliptical distribution and also overestimate correlation. With SSh correlation is estimated better but the shape is still too close to elliptical. MBCn is better able to represent the inter-variable dependence, but the complex structure is captured well only by OA.

The dependence structure between T2m in De Bilt and De Kooy is not well captured compared to observations (grey) by the raw ECMWF forecast (blue, Fig. 6), likely due to the coastal location of De Kooy. The EMOS-based methods (EMOS-ppc, ECC, SSh) generally correct the bias but the forecasts are too cold compared to the observations. Only MBCn and OA are able to realistically forecast the relationship between temperature in De Bilt and De Kooy in the lower tail of the distribution, and only OA is able to correctly forecast the relationship between stations on warm days (Fig. 6).

The variogram scores for T2m and DPT across stations supports these results (Fig. 7). The mean variogram score for T2m and DPT shows how well the forecast methods are able to simulate the spatial dependence. EMOS-ppc (without reshuffling) does not improve on the raw forecast

after the first lead time. ECC is best able to forecast the spatial dependence structure at the +12 and +36 h lead times, while MBCn is most skilful at moderate lead times (+84 to +132 h), and OA is the most skilful at the longest lead times (+180 to +228 h). There are fewer differences between the most skilful methods (ECC, MBCn, OA) for DPT.

The energy score for both variables and all stations shows few differences between the methods (Fig. 8). The raw ensemble is substantially less skilful at short lead times but is comparable to EMOS-ppc (without reshuffling) at the longer lead times. It is unsurprising that there are few differences in the energy score, as it is known to be rather insensitive to errors in the dependence structure (Scheuerer and Hamill, 2015b).

Similarly to the variogram score for T2m, the variogram score for the inter-variable and spatial dependencies shows that ECC is the most skilful at shorter lead times, while MBCn and OA are the most skilful at moderate and longer lead times (Fig. 9). This is possibly due to a lack of calibration at short lead-times in MBCn and OA. The SSh is the least skilful of the methods that can correct the dependence structure, but is of course more skilful than the raw ensemble and the unshuffled EMOS-ppc. This is possibly due to the selection of sub-optimal dates in the dependence template due to the selection of a relatively wide window ( $\pm 5$  days) for a variable like temperature with such a strong seasonal cycle. An initial sensitivity analysis with a wider  $\pm 10$  day window resulted in slightly less skilful forecasts, suggesting that this could be the case. Future work could examine the value of more advanced methods to select dates, such as the minimum-divergence Schaake Shuffle (Scheuerer et al., 2017). The poor performance of the SSh compared to ECC is consistent with Wilks (2015) in the case where the sample size is the same. This suggests that flow dependent error characteristics are important for The Netherlands and that the ECMWF model does a reasonable job at simulating the dependence structure, despite an

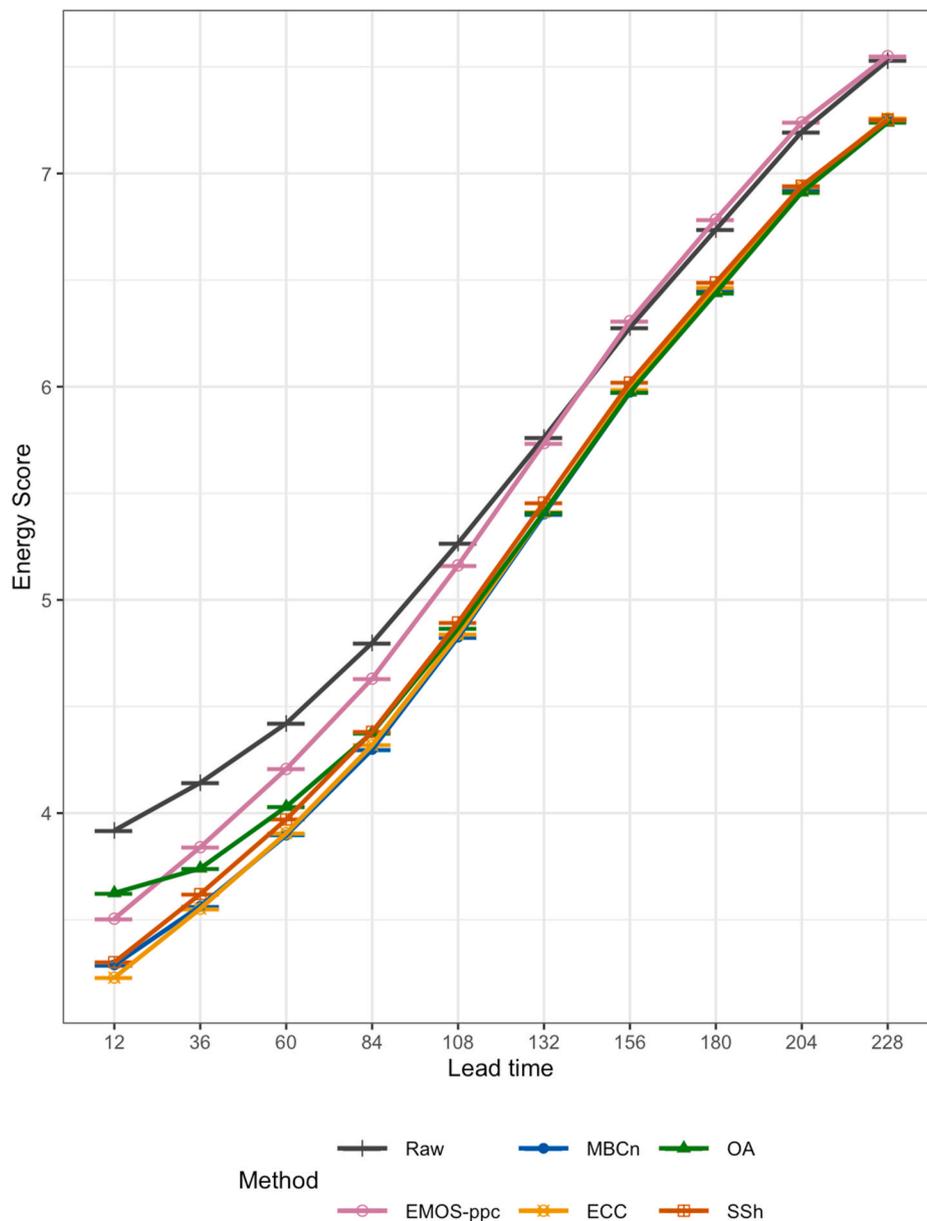


Fig. 8. The mean energy score of spatial T2m-DPT forecasts (all stations).

over-estimation of the inter-variable correlation (Fig. 1).

Regarding the heat index, we first verify the forecast skill at each station using the CRPS (Fig. 10). All methods except for OA improved the forecast at the +12 h lead time compared to the raw ECMWF forecast. De Kooy (station number 235) is poorly forecast by all methods except OA. The bias in the raw ensemble is very large and it is reduced by all methods but remains substantially larger than at the other stations for most methods (EMOS-ppc, SSh, ECC and MBCn), while it is comparable to the other stations in OA. At longer lead times (+132 and +228 h) all methods improve upon the raw forecast, with fewer differences between the methods, particularly at the +228 h lead time.

The variogram score across stations for the heat index shows the largest differences between methods (Fig. 11). There is a substantial decrease in the variogram score for MBCn and OA compared to all other methods. Indeed, from the +108 h lead time there is a significant reduction in the variogram score compared to ECC.

#### 4. Discussion and conclusions

We have compared two novel multivariate post-processing methods, MBCn (Cannon, 2018) and OA in the spirit of Chemozhukov et al. (2017) and similar to Robin et al. (2019). MBCn and the method of Robin et al. (2019) have been used previously to adjust biases in climate model simulations but not yet in a forecasting context. EMOS is the benchmark method for calibration of univariate ensemble forecasts, but the reshuffling methods that are required for the re-introduction of the dependence structure for multivariate outcomes both have drawbacks. In ECC it is assumed that the raw ensemble is able to simulate the dependence structure well, and in SSh the selection of the template from observations is important.

We demonstrate that while the multivariate quantile mapping methods are less skilful than ECC at short lead times, they are more skilful at longer lead times in terms of the variogram score, which measures how well dependencies are represented. Future work should assess how these methods compare with other verification metrics, such as the Dawid–Sebastiani score (Wilks, 2020). EMOS is the benchmark

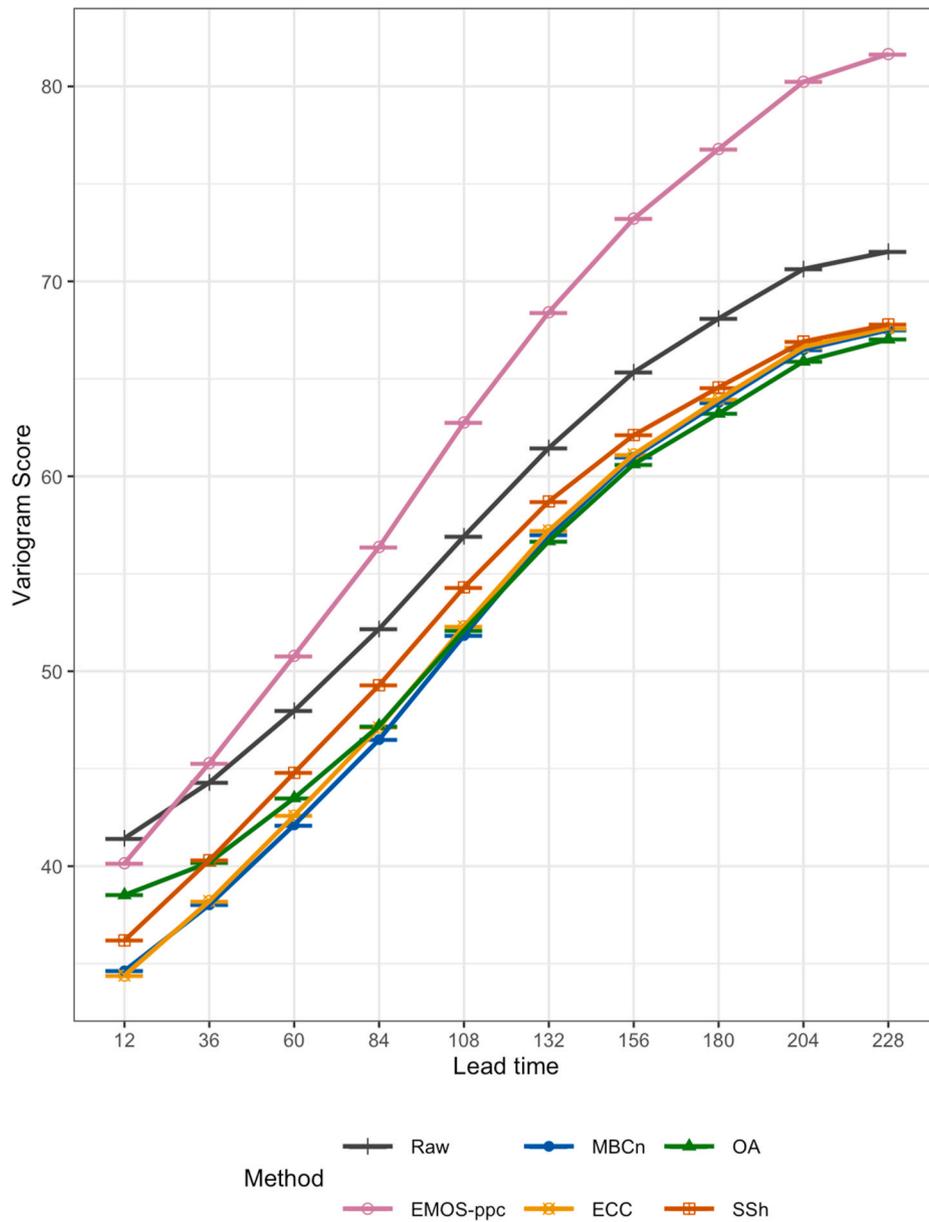


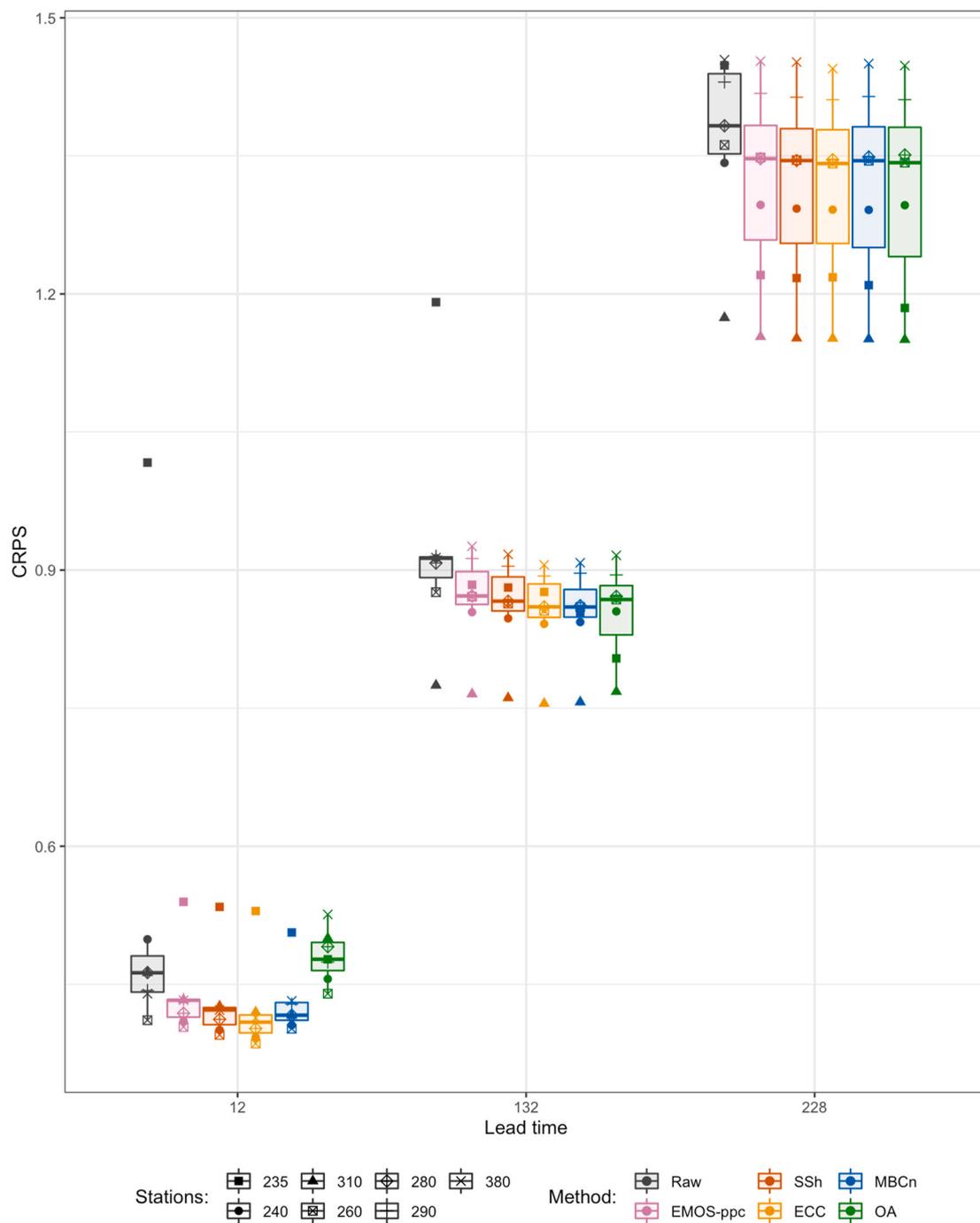
Fig. 9. The mean variogram score ( $p = 0.5$ ) of spatial T2m-DPT forecasts (all stations).

method and it is very hard to improve upon it for temperature, as temperature adheres very well to the Gaussian distributional assumption. For other variables, like precipitation or wind speed, there is less certainty about which forecast distribution to assume, so MBCn or OA could be even more successful for these other variables in the context of multivariate bias correction, however this remains an open question and further research into these methods with other variables is required. MBCn and OA make two changes compared to ECC, as they simultaneously relax the marginal assumption and define the dependence structure differently. Application of these new methods to other variables, with a more uncertain distribution, and comparison with benchmark methods that relax the marginal assumptions can help to disentangle the influence of these two changes.

We have adjusted each member separately in MBCn and OA for computational reasons. This means that for these methods we have not calibrated the ensemble, and the spread has not been adjusted as it is in EMOS. This is one drawback of the current method that could be improved in future applications. However, the lack of spread adjustment penalises the quantile mapping based methods more than the EMOS

methods, and so we tend to underestimate their potential skill. Additionally, it is interesting to see the influence of this choice on the results. It is likely the reason why ECC is most skilful at shorter lead times, when it is well-known that the raw ensemble is under-dispersed and requires the largest adjustment to the spread. Somewhat surprisingly, it also implies there is little to be gained (in terms of the variogram score) by adjusting the spread after the second forecast day.

Further research is needed into multivariate post-processing methods. EMOS, and other more advanced machine learning methods like random forests, are able to use information from a number of other potential predictor variables. For example, in EMOS for temperature we could allow the mean and/or standard deviation of the forecast distribution to depend on variables such as cloud cover, or indices of large-scale circulation (Velthoen et al., 2020). These other potential predictors could be valuable in situations where the relationship between observed temperature and the ensemble mean forecast temperature is not strong (e.g. in De Kooy in our application). One drawback of the quantile mapping methods is that they are not able to use this additional information. Other bivariate parametric methods, such as fitting a



**Fig. 10.** The CRPS of the heat index forecasts at the +12, +132 and +228 h lead times. The colours indicate the method and the boxplots show the distribution of the CRPS over the stations, with the line indicating the median, the boxes showing the interquartile range and the whiskers showing  $1.5 \times$  the interquartile range. Shapes indicate the mean CRPS of individual stations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

bivariate Gaussian distribution could be tried, as they have the benefit of being able to include additional predictor information and model the dependencies well. However they could not capture the non-Gaussian dependence structures seen here. Furthermore, it is likely that they would not be feasible in a high-dimensional case. It would be interesting in the future to quantify the improvement in forecast skill from using additional potential predictors in a more advance machine learning method with the quantile mapping based approaches. Another avenue for future research is to additionally correct for dependencies across time. There are bias adjustment approaches that can deal with very high dimensionality (Vrac, 2018), and can adjust dependencies in space and

time concurrently. The temporal order of events can be highly relevant for impacts (Zscheischler et al., 2020).

In conclusion, we have shown the value of multivariate bias-correction approaches in a weather forecasting context. Our results highlight the importance of accounting for spatial and inter-variable dependencies in statistical processing of ensemble forecasts. This might be of particular relevance for hazard indices that are computed from correlated meteorological variables, or for emergency planning of spatially correlated hazards. Our approach thus paves the way for improved early warnings for multivariate hazards.

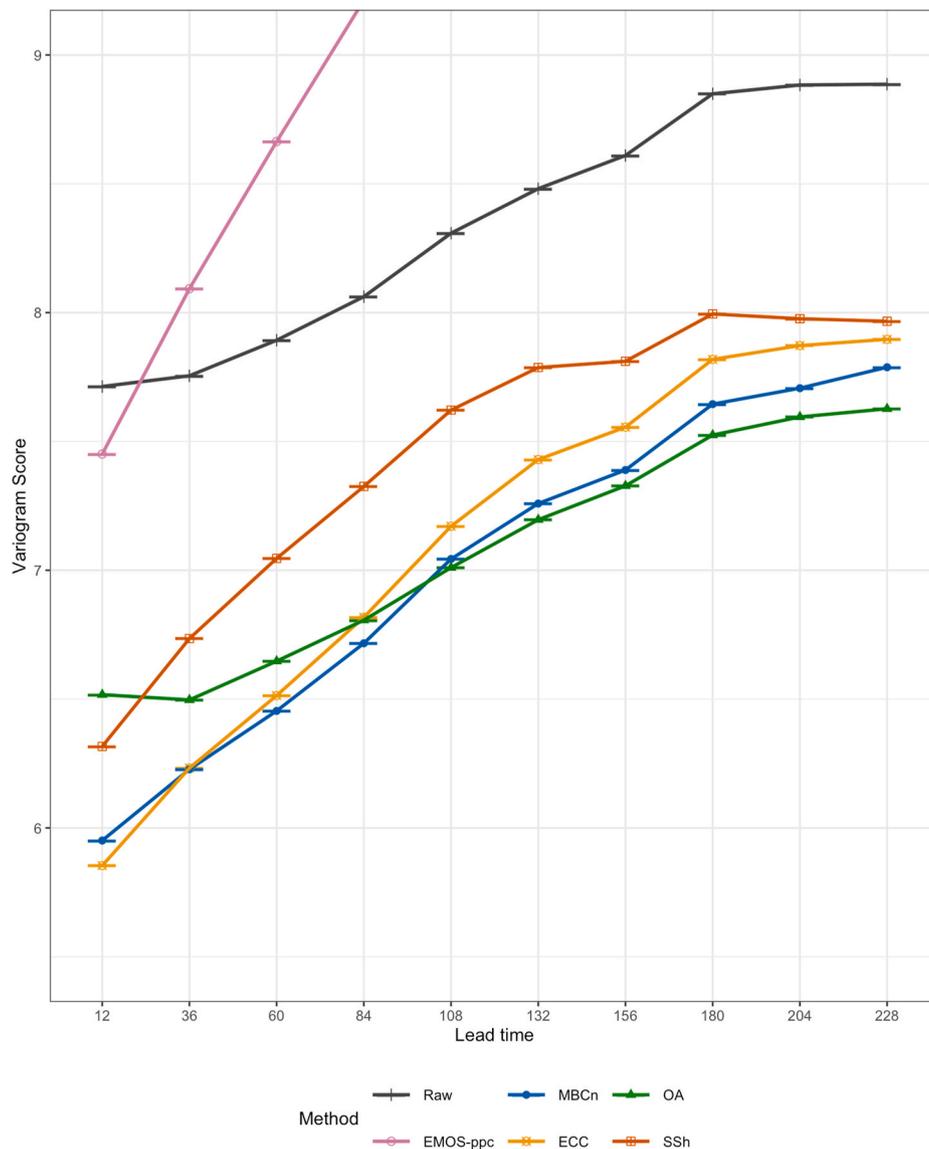


Fig. 11. The mean variogram score ( $p = 0.5$ ) of the spatial heat index forecasts (all stations). The variogram score for EMOS-ppc extends off the top of the figure.

**CRedit authorship contribution statement**

**Kirien Whan:** Conceptualization, Formal analysis, Investigation, Writing - original draft, Funding acquisition. **Jakob Zscheischler:** Conceptualization, Writing - review & editing. **Alexander I. Jordan:** Methodology, Writing - review & editing. **Johanna F. Ziegel:** Methodology, Writing - review & editing.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgments**

This project was supported by a Short-Term Scientific Mission from the European COST Action DAMOCLES (CA17109). Jakob Zscheischler acknowledges funding from the Swiss National Science Foundation (Ambizione grant 179876) and the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, Grant Agreement VH-NG-1537). We thank Maurice Schmeits (KNMI), Kees Kok (KNMI) and Kate Saunders (TU Delft, KNMI) for useful discussions. We are grateful to the anonymous reviewers for their comments, which have helped to improve the manuscript.

**Appendix A. Correlations between T2m and DPT**

The increase in the correlation between T2m and DPT at each lead time in the raw ECMWF forecasts from the observations are shown below.

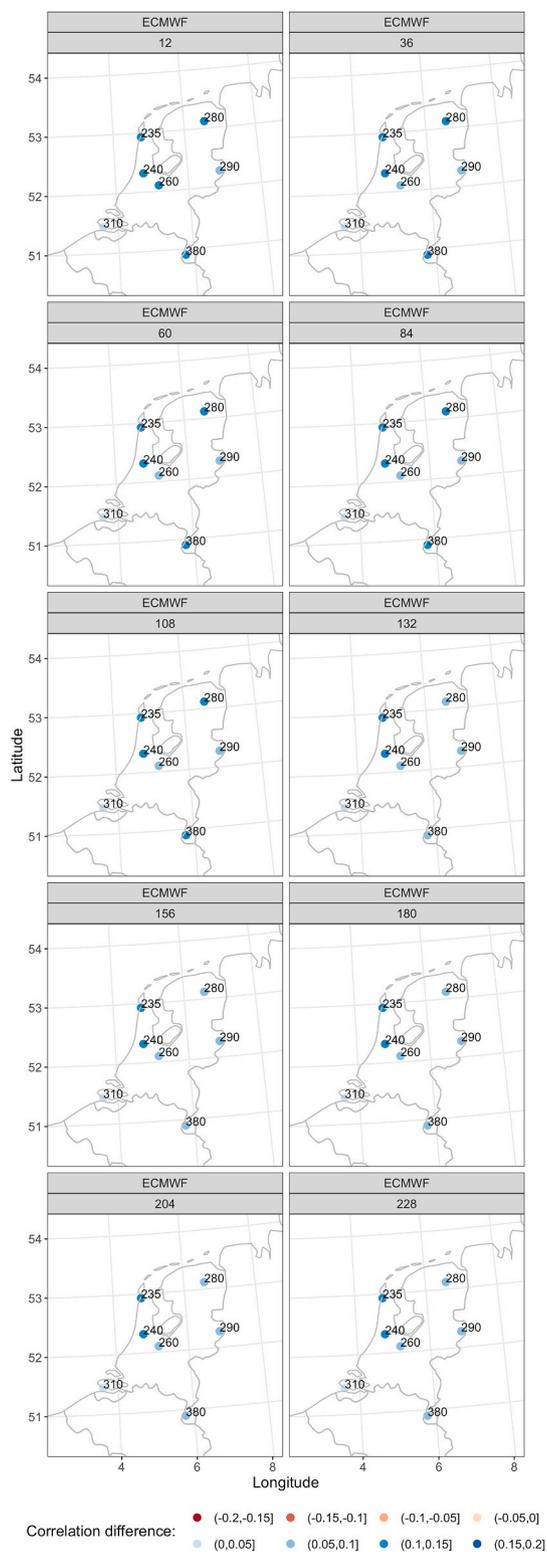


Fig. A.12. The difference in the correlation between T2m and DPT in all members of the ECMWF ensemble and the observations. Numbers indicate the station identifiers that can be seen in Table 1.

**Appendix B. Wet-bulb globe temperature (WBGT)**

We calculate the WBGT index according to Lemke and Kjellstrom (2012), following equations B.1 to B.4.

$$WBGT = 3.94 + 0.567T + 0.393 \left\{ 0.06105RH \exp \left( \frac{17.27T}{237.7 + T} \right) \right\} \tag{B.1}$$

$$ewT = 6.1121 * \exp\left\{\left(\frac{17.502T}{240.97T}\right)\right\} \quad (\text{B.2})$$

$$ewTd = 6.1121 * \exp\left\{\left(\frac{17.502Td}{240.97Td}\right)\right\} \quad (\text{B.3})$$

$$RH = \left(\frac{ewTd}{ewT}\right) * 100 \quad (\text{B.4})$$

## References

- Cannon, A.J., 2018. Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables. *Clim. Dynam.* 50, 31–49.
- Chernozhukov, V., Galichon, A., Hallin, M., Henry, M., 2017. Monge–Kantorovich depth, quantiles, ranks and signs. *Ann. Stat.* 45, 223–256.
- Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R., 2004. The schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* 5, 243–262.
- François, B., Vrac, M., Cannon, A.J., Robin, Y., Allard, D., 2020. Multivariate bias corrections of climate simulations: which benefits for which losses? *Earth Syst. Dynam.* 11, 537–562. <https://doi.org/10.5194/esd-11-537-2020>.
- Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Mon. Weather Rev.* 133, 1098–1118.
- Gneiting, T., Stanberry, L.L., Gritti, E.P., Held, L., Johnson, N.A., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* 17, 211.
- Hopson, T.M., Webster, P.J., 2010. A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: forecasting severe floods of 2003–07. *J. Hydrometeorol.* 11, 618–641.
- Hornik, K., 2005. A CLUE for CLUster ensembles. *J. Stat. Software* 14. <https://doi.org/10.18637/jss.v014.i12>.
- Hornik, K., 2019. clue: cluster ensembles. URL <https://CRAN.R-project.org/package=clue>. r package version 0.3-57.
- Jordan, A., Krüger, F., Lerch, S., 2019. Evaluating probabilistic forecasts with scoringRules. *J. Stat. Software* 90, 1–37. <https://doi.org/10.18637/jss.v090.i12>.
- Joslyn, S.L., LeClerc, J.E., 2012. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.* 18, 126.
- Keune, J., Ohlwein, C., Hense, A., 2014. Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Mon. Weather Rev.* 142, 4074–4090.
- Lang, M.N., Lerch, S., Mayr, G.J., Simon, T., Stauffer, R., Zeileis, A., 2019. Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Proc. Geophys. Discuss.* 1–18.
- Lemke, B., Kjellstrom, T., 2012. Calculating workplace wbgt from meteorological data: a tool for climate change assessment. *Ind. Health* 50, 267–278.
- Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., Stafford-Smith, M., 2014. A compound event framework for understanding extreme impacts. *Wiley Interdiscip. Rev.: Clim. Change* 5, 113–128.
- Lerch, S., Thorarindottir, T.L., 2013. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus Dyn. Meteorol. Oceanogr.* 65, 21206.
- Maraun, D., 2013. Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. *J. Clim.* 26, 2137–2143.
- Palmer, T., Shutts, G., Hagedorn, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., 2005. Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet Sci.* 33, 163–193.
- Pinson, P., 2012. Adaptive calibration of (u, v)-wind ensemble forecasts. *Q. J. R. Meteorol. Soc.* 138, 1273–1284.
- Pitie, F., Kokaram, A.C., Dahyot, R., 2005. N-dimensional probability density function transfer and its application to color transfer. In: Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 1. IEEE, pp. 1434–1439.
- Pitié, F., Kokaram, A.C., Dahyot, R., 2007. Automated colour grading using colour distribution transfer. *Comput. Vis. Image Understand.* 107, 123–137.
- Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape (with discussion). *Applied Stat.* 54, 507–554.
- Robin, Y., Vrac, M., Naveau, P., Yiou, P., 2019. Multivariate stochastic bias corrections with optimal transport. *Hydrol. Earth Syst. Sci.* 23, 773–786.
- Saunders, K., Whan, K., 2020. depPPR: depPPR - dependence Post-processing in R. URL <https://github.com/katerobsau/depPPR>. r package version 0.1.0.
- Schefzik, R., Thorarindottir, T.L., Gneiting, T., et al., 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.* 28, 616–640.
- Scheuerer, M., Hamill, T.M., 2015a. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* 143, 4578–4596.
- Scheuerer, M., Hamill, T.M., 2015b. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Weather Rev.* 143, 1321–1334.
- Scheuerer, M., Hamill, T.M., Whitin, B., He, M., Henkel, A., 2017. A method for preferential selection of dates in the schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.* 53, 3029–3046.
- van Straaten, C., Whan, K., Schmeits, M., 2018. Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *J. Hydrometeorol.* 19, 1815–1833.
- Velthoen, J., Cai, J.J., Jongbloed, G., 2020. Interpretable random forest models through forward variable selection. *J. Appl. Stat.* (under review) URL <https://arxiv.org/abs/2005.05113>.
- Voisin, N., Schaake, J.C., Lettenmaier, D.P., 2010. Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Weather Forecast.* 25, 1603–1627.
- Vrac, M., 2018. Multivariate bias adjustment of high-dimensional climate simulations: the Rank Resampling for Distributions and Dependences (R<sup>2</sup>D<sup>2</sup>) bias correction. *Hydrol. Earth Syst. Sci.* 22, 3175–3196. <https://doi.org/10.5194/hess-22-3175-2018>.
- Whan, K., Schmeits, M., 2018. Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Mon. Weather Rev.* 146, 3651–3673.
- Wilks, D.S., 2015. Multivariate ensemble model output statistics using empirical copulas. *Q. J. R. Meteorol. Soc.* 141, 945–952.
- Wilks, D.S., 2020. Regularized dawid–sebastiani score for multivariate ensemble forecasts. *Q. J. R. Meteorol. Soc.* 146 (730), 2421–2431. <https://doi.org/10.1002/qj.3800>.
- Zscheischler, J., Fischer, E.M., Lange, S., 2019. The effect of univariate bias adjustment on multivariate hazard estimates. *Earth Syst. Dynam.* 10, 31.
- Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R.M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M.D., Maraun, D., Ramos, A. M., Ridder, N., Thiery, W., Vignotto, E., 2020. A typology of compound weather and climate events. *Nat. Rev. Earth and Environ.* 1, 333–347. <https://doi.org/10.1038/s43017-020-0060-z>.
- Zscheischler, J., Westra, S., Van Den Hurk, B.J., Seneviratne, S.I., Ward, P.J., Pitman, A., AghaKouchak, A., Bresch, D.N., Leonard, M., Wahl, T., et al., 2018. Future climate risk from compound events. *Nat. Clim. Change* 8, 469–477.