List of relevant changes in manuscript

- Eq. (1) and Eq. (2) were changed.
- Improved explanations of linear scaling and quantile mapping methods in Sect. 2.3.1.
- Sect. 2.4.3. was extended to include an explanation of the Anderson-Darling (AD) test.
- Sect. 3.1.2 was edited to include changes in Fig. 3.
- We included the discussion of the effect of varying ensemble member size on the CRPSS in Sect. 3.1.2.
- Section 3.1.3. was heavily edited to include results of one grid point only, as per reviewer suggestion.
- Section 3.2.3. was extended with the discussion of the results of the Anderson-Darling (AD) test.
- Sect. 8. (Figure Captions) was removed.
- Appendix A was removed.

Changes in Figures.

- Figs. 1, 2, 3, 4, 5, 9, 10, 11 (in the revised version, 12). Layout was improved to comply with reviewers suggestions.
- Fig. 6. The PIT diagram includes only the results for one grid point.
- We added a figure with the reults of the AD test for uniformity of the PIT diagrams (Fig. 11).
- Fig. 12 was removed.

Supplement

- Figures were improved as per reviewers suggestions.

We appreciate the time spent carefully reviewing this manuscript. We are certain that your comments and suggestions will improve the quality of this paper.

Anonymous Referee #1

This paper investigates the performance of precipitation and temperature forecasts from ECMWF System 4, as well as derived reference evapotranspiration. The authors also look at the impact of two simple postprocessing methods: linear scaling and quantile mapping, on the performance of these forecasts. Raw forecasts tended to be overconfident, and regions with biases also corresponded to regions with lower skill. Both linear scaling and quantile mapping performed well in removing biases, quantile mapping was better suited to improve statistical consistency. General comment I found the paper well-written and I think that it provides both didactic explanations for the methodology as well as an in-depth and comprehensive analysis of the skill and bias patterns. It also fits in the Subseasonal-to-seasonal special issue since it follows and complements nicely the paper by Crochemore et al. (2016) published in this same issue.

I list below a few comments about this version of the paper. These comments are mainly technical. My main remark would be that the figures are generally too small and thus difficult for the reader to read and interpret. I detail this point further down.

Major comments and general questions

Section 2.2: What do you think is the impact of the interpolation method on the results and areas with skill? How common is the inverse distance weighting to interpolate meteorological variables? Couldn't this induce yet another bias in the forecasts? If observations had been upscaled to preserve the scale of the GCM forecasts, would you expect similar results?

We made the decision of the interpolation procedure based on its simplicity and on studies that have used a similar method (Wood et. al., 2002; Voision et al., 2010; Tian, 2014). We are aware of the issues that come up when using such an interpolation method. However, the disadvantages of this choice in comparison to others (spatial interpolation, upscale observations, statistical downscaling) is a study of its own and out of the scope of this manuscript. ~~Besides, computing aerial precipitation it is not enough for our purposes.~~Moreover, ~~We~~ we needed to investigate the behavior of the forecasts on a resolution relevant to hydrological forecasting. Additionally, we do not believe that, at least for monthly accumulated values, values of skill in accuracy will change in a significant manner. Evidence for the latter claim can be found in Fig. 1 below where, as you suggest, the comparison is done on the 70 km grid box versus inside box average.

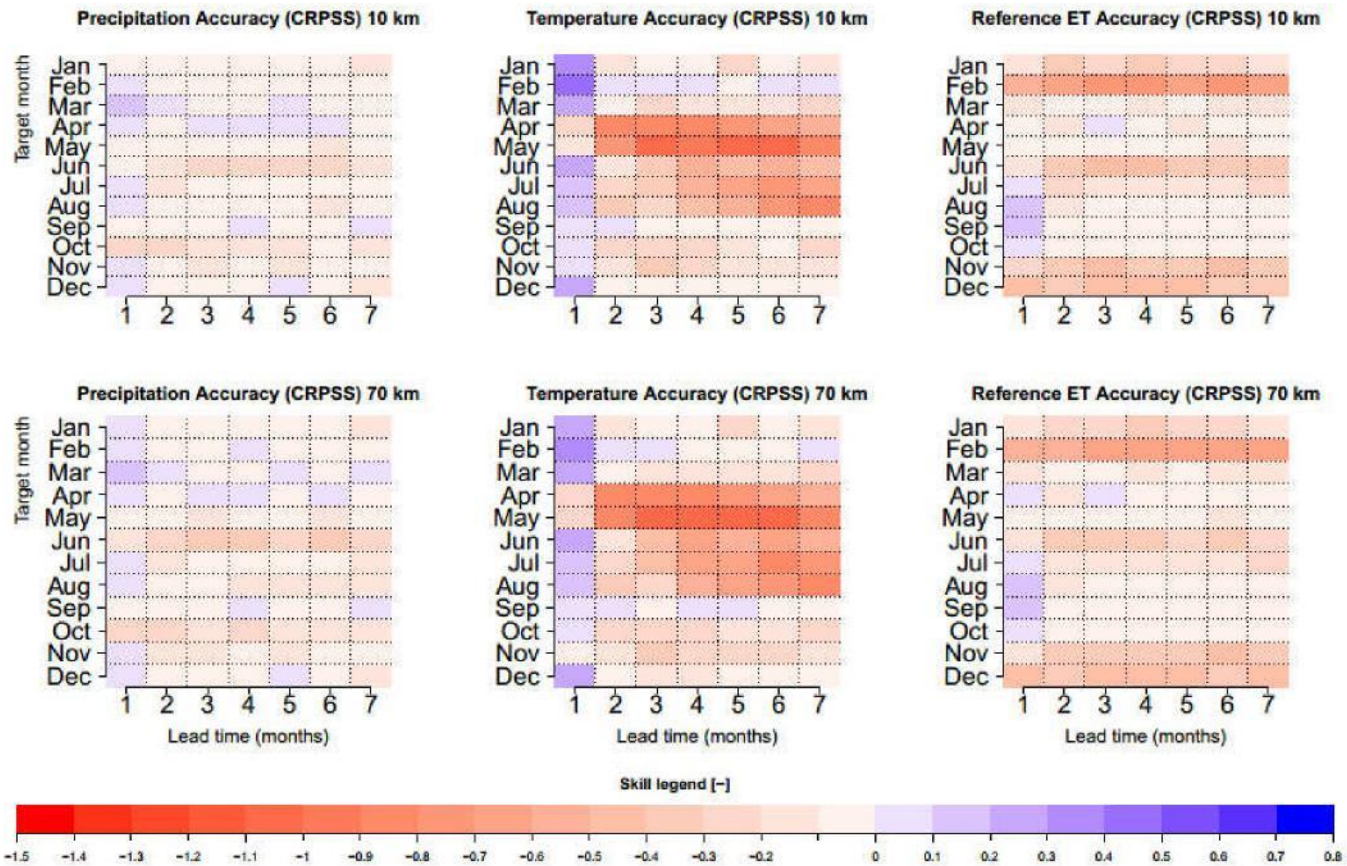


Fig 1. Skill in terms of accuracy of monthly values of raw forecasts at the 10 km grid (first line) and 70 km grid (second line). Y-axis represents the target month and X-axis represents the different lead times at which target month is forecasted.

Section 3.2.4 and Figure 11: It was unclear to me why linear scaling impacted the number of dry days. If dry days are defined as days with no precipitation, and linear scaling solely consists in applying a multiplicative factor, the number of zero-values should not be affected. Did you define a threshold to determine dry days? Please clarify this.

We failed to note in Sect. 2.3.1 that a previous step was introduced before applying the correction factor for precipitation. An analysis was made as to the threshold for which the number of dry days in a given month was equal to the observed number of dry days. On average along Denmark, this value is of 1.5 mm/day. That is why in Sect. 3.2.4 and in Fig. 11, LS seems to increase the skill in predicting the number of dry days. It is then a consequence of this threshold rather than the method itself. Note that this previous step was not introduced in QM as it will map small forecast precipitation values with larger observed values. We will make this clearer in the new version of the manuscript.

Technical comments

P2 L7: Replace “Despite of the efforts” by “Despite the efforts”.

Yes, will do.

P2 135: I suggest adding "only" after "for precipitation", for clarification.

We will change this accordingly.

P4 12-4: Please check the indices in equations 1 and 2. It seems that index i is used to represent different things: the year for which the correction is applied and a sum that runs from 1 to N while excluding the year for correction itself (previously represented by i).

You are correct. Thanks for noticing this, we will correct it.

P4 16: If I understand correctly, N is equal to the numbers you have, i.e. 24 years from 1990 to 2013. Is that correct?

N is equal to the number of years. The summation runs over N-1 years as the year which is corrected is withdrawn from the sample. ~~As the previous notation, it will be N-1.~~

P8 135: Replace "thorough" with "through".

P9 16-8: Could you please clarify this point?

There are two features at play that might prevent us from detecting biases in ensemble spread with the presented configuration of the PIT diagrams:

(1) The spatial pooling of the zi.

(2) The monthly accumulation of the variables as stated in P9 L6-8.

In the updated version of the manuscript we will address the effect of each feature to clarify this point as also suggested by Reviewer # 3 in comment 3a.

P9 111: Replace "The second and third columns column" by "The second and third

rows" Yes, will do.

P10 127-29: "This fact may be ... in the raw forecasts (Fig. 7)" could you please reformulate this

sentence? Yes, will do.

P11 113: It seems that in this context and given the following sentence, "sharpness" can hardly be an advantage.

We will reformulate the sentence accordingly.

P11 121: I suggest replacing "act equally good" by "perform equally well".

Yes, will do.

P11 127-30: Please reformulate these sentences.

Yes, will do.

P12 13: Replace "The second is that the exclusion" with "The second is the exclusion".

We are aware of the issue with the plots, we will improve the layout of all figures as you carefully suggested.

Figures 1, 4 and 6 and maps in the Supplement: The maps are too small to easily distinguish the patterns. In addition, it is difficult to spot the stars in Figure 6, both due to the size and the colors. I suggest making the maps bigger, and if necessary, changing the color of the stars in Figure 6.

Figure 2: Please explain the x-axis somewhere or make the years fully explicit.

Figure 3: The x-axis is not the same size in all three graphs. The size used in the left-hand graph is easier to read.

Figure 5: Please increase the size of the axis labels. Consider replacing “lt” by “lead times”. Please also reformulate the last sentence in the legend.

Figure 9 and similar graphs in the Supplement: Please also increase the labels here.

Figure 10: I recommend moving the legend to the first or second graph for readability.

Figure A12: I think N(0.0.3) should be N(0,0.3).

Reviewer # 2

The paper is well written and the results could be of interest for researchers testing extended range and seasonal forecasts especially for hydrological applications. Therefore the paper is worth to be published after some minor corrections and/or inclusions of additional explanations.

1. On page 3 you describe the ECMWF forecast data. I was wondering why there some months with 15 and some with 51 members? Could you explain this?

Did you include some corrections in the skill scores for the 15 ensembles (e.g. Müller, W.A., C. Appenzeller, F.J. Doblas-Reyes, and M.A. Liniger, 2005: A Debiased Ranked Probability Skill Score to Evaluate Probabilistic Ensemble Forecasts with Small Ensemble Sizes. J. Climate, 18, 1513–1523, https://doi.org/10.1175/JCLI3361.1) in order to make them comparable with the 51 members?

Have there been any model changes within these 24 years? If yes, has this been taken into consideration?

First, this is the data we received from the meteorological forecast provider (ECMWF). We assume that the increase of ensemble size for February, May, August and November is done with the objective of increasing quality of the forecasts of the upcoming season, for example summer (JJA) for forecasts initialized in May. We will clarify if this is the case in the updated version of the manuscript.

Second, we did not include corrections to the estimated CRPS as in Ferro et al., (2008) and references therein. We will ~~evaluate whether this correction is applicable to our case and~~ apply the correction to obtain unbiased estimator of CRPS as done in Crochemore et al., (2016). ~~If we, however, consider that the assumptions of such estimators are not met (i.e., perfect reliability) we will explain our decision and discuss it the new version of the manuscript.~~

Finally, to the knowledge of the authors, there was no model update in the 24 year period used in the present manuscript.

2. Also on page 3 line 34 you give the dimension of 662x12x7x24. Does this mean grid cells x months x forecasts x years ? Could you please clarify this?

It means grid cells, months, lead times, years. This will be clarified in the revised manuscript.

3. On page 4 you explain the the QM approach. In line 16 you write that the EDF has been trained. I would rather call this process fitting.

Yes, you are correct. We will correct the manuscript accordingly.

4. Regarding the skill (from page 5 onwards): The CRPS is a global score combining the reliability and sharpness aspects. You mention that you use the CRPS as a general measure of accuracy, but you don't say why it is general. I think this should be included for readers who are not familiar with the CRPS (e.g. Gneiting, T., A. Raftery, A. Westveld III, and T. Goldman (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133(5), 1098–1118)

We will improve our explanation of the CRPS for better clarity in the next version of the manuscript.

5. Have you tried to eliminate the drizzle effects and include different thresholds for zero precipitation? This could be interesting for analysing the dry periods.

We failed to note in Sect. 2.3.1 that a previous step was introduced before applying the correction factor for precipitation. An analysis was made as to the threshold for which the number of dry days in a given month was equal to the observed number of dry days. On average along Denmark, this value is of 1.5 mm/day. That is why in Sect. 3.2.4 and in Fig. 11, LS seems to increase the skill in predicting the number of dry days. It is then a consequence of this threshold rather than the method itself. Note that this previous step was not introduced in QM as it will map small forecast precipitation values with larger observed values. We will make this clearer in the new version of the manuscript.

6. I agree with Reviewer 1 that the Figures are difficult to read. Regarding Figure 2 it would be interesting to compare the boxplots of the forecasts with boxplots of the climatology in order to see the median, interquartile range.

We will improve the layout of all figures in the updated version of the manuscript. In Fig. 2 you can see the boxplots of the forecasts (black) and the boxplots of climatology (light blue) as you suggest.

7. I don't think that the PIT diagram has to be explained in that detail and you could delete Appendix A and Figure A12. You can find the same Figures in Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11(4), 1267–1277 and in Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45, W00B14, doi:10.1029/2008WR006825.

We will remove the Appendix and Figure A12 and refer to the papers you mention instead.

Reviewer # 3

1 General comments

The paper by Lucatero et al. describes an assessment of uncorrected and postprocessed GCM forecast over Denmark during the period 1990-2013. The study adresses a critical issue for GCM forecast users, especially in the field of hydrology where uncorrected forecast often lacks sufficient skill to be used as input for hydrological applications. Most methods applied in the paper are sound, for example the use of two well established post-processing techniques and a leave-out cross validation scheme. The paper is well written with a clear and concise structure. However, we believe that there is scope to improve its content before it can be published. Our major comments are listed below:

1. The approach that was adopted by the authors to downscale the ECMWF forecast is questionable. The authors applied an inverse weighting algorithm to convert the 70km resolution ECMWF grid to a 10 km resolution, and then used the downscaled data to post-process and analyse forecast performance. By doing this, they smooth ECMWF rainfall surfaces and break the conservation of mass, which artificially reduces the skill of uncorrected forecasts. To circumvent this problem, we recommend performing the analysis undertaken by the authors at the resolution of ECMWF forecasts (i.e. 70km), using a simple aggregation method for gridded observation data (see for example Schepen et al. (2014)). This alternative approach would eliminate the need for a downscaling algorithm, and provide a direct assessment of uncorrected ECMWF forecasts compared to post-processed forecasts. We understand the value of downscaling to work at a meaningful scale for hydrological applications. However, downscaling is a research topic in its own, and its impact should not mask the skill of the uncorrected forecasts.

Without a proper assessment of uncorrected forecasts, it is difficult to select the appropriate downscaling

model. See the reply to reviewer #1 on the same topic.

2. The quality and resolution of several figures is clearly below the standard of an international scientific journal. We strongly recommend redrawing figures 1, 4 and 6, increasing the resolution and/or converting them to a vector format. Unfortunately, with such low figure quality, it becomes difficult to check the comments made by the authors in reference to those figures. Additional comments on the figures are provided in the next section to improve their readability.

We are aware of this issue, we will improve the layout of all figures.

3. The analysis of forecast reliability (or statistical consistency as per the author's nomenclature) lacks important information to properly assess forecast performance:

- **It is not clear which variables are used to draw the PIT plots (figure 7). Such plots require a single series of PIT values computed from matched pairs of observations and forecasts. The authors do not precise if the observations and forecasts are coming from a single grid cell, or from a spatial aggregation (e.g. the whole Denmark). This point is important to understand their difficulties in interpreting the PIT plot (see Line 6 page 9:"issues (...) are not remarkably clear"). We suggest drawing the PIT plots for a selected set of grid points and one lead time representing the main characteristics of the PITs across the study area.**

Thanks for noticing this point. We compute the z_i's (Page 6 Line 7) of N pairs of forecast-observations of each grid cell which then ends up with 662 x N values of z_i's for the whole Denmark (for P, 724 for T and ET0). Then for the PIT

diagram we pool all 662 x N z_i's. We are aware that by doing this we are pooling points that might not be independent which perhaps has an influence in limiting our ability to detect biases in spread (Hamil, 2000).

We then will replace Fig. 7 as well as Fig. S7 to Fig. S9 in the supplement with PIT diagrams ~~of a selection of grid points~~for a grid point as you suggest. We believe that this will reflect the biases in the mean observed in Fig. 4 and Fig. S1 to Fig. S3 in the supplement.

- **The analysis of reliability lacks a quantitative criterion similar to the skill scores. A standard approach is to compute the pvalue of a uniformity test such as the Kolmogorov-Smirnov test (Laio and Tamea, 2007). However, instead of the KS test, we recommend the Anderson-Darling test (Anderson and Darling, 1952), which exhibits a greater power (Noceti et al., 2003) and better ability to detect deviations from uniformity of the extremes. We strongly suggest computing the p-value of such tests for all grid points of the domain and all lead times, and then summarise the results by counting how many cells pass the test with a 5% threshold for a given lead time. This will provide a consistent and reproducible assessment of forecast reliability across the study area.**

We appreciate the suggestion and believe that will add more quantitative evidence on the discussion of statistical consistency. We will make use of the AD test for uniformity and create a figure similar to Fig. 9 for statistical consistency. The results will be then discussed in the updated version of the manuscript.

4. Temperature is treated differently than rainfall and PET both in the computation of the bias score and in the implementation of the LS post-processing model. This is quite confusing and should be harmonised. We recommend that all bias scores be computed in relative terms to facilitate comparison. The LS model should be applied to the three variables in two modes: multiplicative (which is the configuration used for P and PET)and additive(used for T).This approach would provide a clear advice on the choice of the LS configuration.

The configuration of the LS is as chosen because of the nature of the variables. For example, P and ET0 cannot be negative. Applying the LS as additive as you suggest can end up with negative P if the correction factor in Eq. 2. is negative (in case of underestimation) and larger than the value of the ensemble member to be corrected ($f_{k,i}$ in Eq. 2). Therefore, we have decided to keep the configuration as it is and explain the reason of the differences in implementation between variables in the updated version of the manuscript.

5. It is not clear how extrapolation is undertaken within the QM model. The authors only indicate that ''Extrapolation is then needed to map ensemble values and percentiles that are outside the training range" (see Page 4, Line 21). This is a critical and frequent problem with the QM method which affects extremes, and is then of particular interest in hydrological forecasts. Please detail the extrapolation method.

We decided to go with a simple approach for the fitting of the empirical CDF as it has been documented that an empirical approach leads to better results (Crochemore, et al., 2016). We do recognize that this choice will have effects especially if the focus is on the evaluation of extreme values. Other approaches might be more suitable for such cases (such as fitting an extreme value distribution to extend the empirical distributions as in Wood et al., 2002). However, this is out of the scope of the present paper. We will address the limitations and advantages of our choice in the discussion of the revised manuscript.

Considering that the paper is covering an important topic for seasonal forecasting, but that the number of points to be improved is quite significant, we recommend the paper to be accepted with major revision. Detailed comments are provided in the following section.

2 Specific comments

1. Page 2, Line 26, “The most used methods are linear scaling and quantile mapping”: Zhao et al. (2017) provides an interesting perspective on the limitations of Quantile-Quantile mapping that could be cited here.

We will cite this important paper as you suggest.

2. Page 3, Line 4, “A statistical consistent forecast system has low (or non-existent) bias in both mean and variance”: correct? ok.

OK.

3. Page 3, Line 14, “we make use of the Makkink equation (Hendriks, 2010) that takes as inputs temperature and incoming short-wave solar radiation from ECMWF System 4”: a mention on the quality of radiation forecast would be useful. A full analysis of radiation forecast performance is out of scope here, but perhaps some references can be cited to understand the impact of radiation inputs in ET0 calculation.

We have searched for studies with specific focus on radiation in Europe at the seasonal scale and only found the following HESS Discussion paper in this special edition.

Greuell, W., Franssen, W. H. P., Biemans, H., and Hutjes, R. W. A.: Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2016-603, in review, 2016.

However, the verification focuses on run-off and discharge only.

~~If we find more on this topic they will be cites to address the performance of radiation forecasts.~~

4. Page 3, Line 20, “The time and spatial variations of the variables can be seen in Fig. 1.”: Fig 1 is not readable, so the accuracy of this comment is hard to assess.

As mentioned above, we will improve the readability of all figures presented.

5. Page 3, Line 38, “once a scale factor has been applied”: In the case of T, the factor is not a

scale, but a shift. Thank you for noticing this. We will improve it accordingly.

6. Page 6, Line 2, “a Wilcoxon-Mann-Whitney test was carried out.”: What was the data used to apply the WMW test? Was it applied to all grid cells for a single lead time? Please clarify.

We applied the test considering N forecast-observation pairs of each grid for a single lead time. We will clarify this in the updated version of the manuscript.

7. Page 6, Line 39, “April shows an overestimation that might be due to the ‘drizzle effect’ in a month where dry days are more common.”: It is quite surprising that this “drizzle effect” is not affecting the bias in March and May where the bias shows an opposite trend compared to April. This statement requires more explanations. We suggest showing the number of dry days per month, to confirm that April has the highest proportion.

You are correct, this statement requires additional information to back up our claim. We will include this information in the updated version of the manuscript.

8. Page 11, Line 13, “One advantage ECMWF System 4 has over ensemble climatology is that forecasts are sharper”: The forecast is sharper but it is not clear if it is reliable. In this case, being sharp is not an advantage, but a problem.

We will reformulate the sentence accordingly.

9. Page 11, Line 24, “QM assumes that there is a linear relationship between ensemble mean and observations, assumption may not hold”: We believe that the authors mean LS instead of QM here. Otherwise we do not see why QM would assume such linear relationship.

We meant that when there is a linear relation between forecast and observations, QM performs better as has been demonstrated in Zhao et al. (2017). We will rephrase this statement accordingly.

10. Figure5: The figure is trying to convey too many information at the same time with a confusing choice of symbol sizes (decreasing symbol size for better sharpness skill score) and color schemes. We suggest using the same presentation than Figure 3 and grouping the two figures into a single one that would provide an homogenous overview of forecast performance.

We will separate the figures to make the readability clearer.

11. Figure 6: The scale of the color bar is different between the 3 plots. As a result, it is impossible to compare the skill between the three variables.

We will set the same color bar for the three variables in question.

12. Figure 8: This plot is difficult to read because the legend does not explain that each location has a different symbol and all curves are shown with the same color. We suggest splitting each one of the three plots into 4 different subplots showing the results for one location only.

We will state the differences more clearly in the figure caption.

References

Crochemore, L., Ramos, M.-H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 20, 3601-3618, https://doi.org/10.5194/hess-20-3601-2016, 2016.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 19–24, doi:10.1002/met.45, 2008.

Hamill, T.M., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.,* **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Tian, D., C.J. Martinez, W.D. Graham, and S. Hwang, 2014: Statistical Downscaling Multimodel Forecasts for Seasonal Precipitation and Surface Temperature over the Southeastern United States. *J. Climate,* **27**, 8384– 8411, https://doi.org/10.1175/JCLI-D-13-00481.1

Voisin, N., J.C. Schaake, and D.P. Lettenmaier, 2010: Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts. *Wea. Forecasting,* **25**,1603–1627, https://doi.org/10.1175/2010WAF2222367.1

Wood, A. W., E. P. Maurer, A. Kumar, and D. Lettenmaier, Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res., 107(D20), 4429, doi:10.1029/2001JD000659, 2002.

Zhao, T., Bennett, J. C., Wang, Q., Schepen, A., Wood, A. W., Robertson, D. E., and Ramos, M.-H. (2017). How suitable is quantile mapping for postprocessing gcm precipitation forecasts? Journal of Climate, 30(9):3185–3196.

On the skill of raw and postprocessed ensemble seasonal meteorological forecasts in Denmark

Diana Lucatero[1], Henrik Madsen[2], Jens C. Refsgaard[3], Jacob Kidmose[3], Karsten H. Jensen[1]

[1]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark
[2] DHI, Hørsholm, Denmark
[3] Geological Survey of Denmark and Greenland (GEUS), Copenhagen, Denmark

Correspondence to: Diana Lucatero (diana.lucatero@ign.ku.dk)

Abstract. This study analyzes the quality of the raw and postprocessed seasonal forecasts of the European ~~European~~ Center of Medium Weather Forecasts (ECMWF) System 4. The focus is given to Denmark located in a region where seasonal forecasting is of special difficulty. The extent to which there are improvements after postprocessing is investigated. We make use of two techniques, namely, linear scaling/delta change (LS) and quantile mapping (QM) to daily bias correct seasonal ensemble predictions of hydrological relevant variables such as precipitation (P), temperature (T) and reference evapotranspiration (E_{T0}). Qualities of importance in this study are the reduction of bias and the improvement in accuracy and sharpness over ensemble climatology. Statistical consistency and its improvement is also examined. Raw forecasts exhibit biases in the mean that have a spatio-temporal variability more pronounced for P and T. This variability is more stable for E_{T0} with a consistent positive bias. Accuracy is higher than ensemble climatology for some months at the first month lead time only and, in general, ECMWF System 4 forecasts tend to be sharper. E_{T0} also exhibits an underdispersion issue, i.e., forecasts are narrower than their true uncertainty level. After correction, reductions in the mean are seen. This however, is not enough to ensure an overall higher level of skill in terms of accuracy although modest improvements are seen for T and E_{T0}, mainly at the first month lead time. QM is better suited to improve statistical consistency of forecasts that exhibit dispersion issues, i.e., when forecasts are consistently overconfident. Furthermore, it also enhances the accuracy of the monthly number of dry days to a higher extent than LS. Caution is advised when applying a multiplicative factor to bias correct variables such as P. It may overestimate the ability that LS has in improving sharpness when a positive bias in the mean exists.

Formatted: Font:

Formatted: Space Before: 6 pt, After: 0 pt, Line spacing: 1,5 lines

Formatted: Font:

Formatted: Font: (Default) Times New Roman, 10 pt

1 Introduction

Seasonal forecasting has gained increasing attention during the last three decades due to high societal impacts of extreme meteorological events that affect a plethora of weather-related sectors such as agriculture, environment, health, transport and energy, and tourism (Dessai and Soares, 2013). Information of weather-related hazards months ahead are important for protection against extremes for these sectors.

General Circulation Models (GCM) have become the state-of-the-art technology for issuing meteorological forecasts at different time scales. GCM-based seasonal forecasting is possible due to signals that can be extracted from slowly changing systems such as the ocean, and to a lesser extent, land, that then translates into a signal in the atmospheric patterns (Weisheimer and Palmer, 2014; Doblas-Reyes et al., 2013). El Nino Southern Oscillation, ENSO, is the strongest of these signals, and its influence on seasonal forecasting is higher near the tropics (Weisheimer and Palmer, 2014).

Seasonal ensemble forecasts have been operational in Europe since the late 1990s provided by the European Center for Medium Range Forecast (ECMWF) (Molteni et al., 2011) and in the U.S. since August 2004 provided by the National Center of Environmental Prediction (Saha et al., 2013). Other examples of operational seasonal forecasts include the ones generated by the Met Office in UK (Maclachlan et al., 2015), the Australian Bureau of Meteorology (Hudson et al., 2013), the Beijing Climate Center (Liu et al., 2015) and the Hydrometeorological Center of Russia (Tolstykh et al., 2014).

ECMWF is a leading center for weather and climate predictions and its seasonal forecasting system is often regarded as the best (Weisheimer and Palmer, 2014). Research on the quality of the atmospheric forecasts has been done for different system versions (Molteni et al., 2011; Weisheimer et al., 2011). The system has also been compared to other GCM (Kim et al., 2012a, 2012b; Doblas-Reyes et al., 2013) or statistical (van Oldenborgh et al., 2005) seasonal forecasting systems.

Despite ~~of~~ the efforts mentioned above and the documented improvements on forecasting skill of meteorological parameters, specially over the tropics (Molteni et al., 2011), several issues still remain. The main one, and specific to forecasting in Europe and North America is that the signal of the main driver of seasonal predictability, the ENSO, has been found to be weak or non-existent (Molteni et al., 2011; Saha et al., 2013) in these regions leading to poor skill of atmospheric variables such as precipitation. For example, Weisheimer and Palmer, (2014) studied the reliability (consistency between the forecasted probabilities and their observed frequencies) and ranked forecasts using five categories from ‘dangerous’ (1) to ‘perfect’ (5) for two regimes of precipitation (wet/dry) and temperature (cold/warm). For the North European region, they found dry (wet) forecasts during summer, started in May to be 'dangerous' ('marginally useful') and dry (wet) forecasts during winter (started in November) to be 'not useful' ('marginally useful'). For temperature, results were less variable among the different categories with winter cold/warm and summer warm forecasts found to be 'marginally useful', and summer cold temperatures forecasts in the category (5) for 'perfect'. Moreover, Molteni et al., (2011) found weak anomaly correlations of precipitation and temperature during the summer for most of the regions located in North Europe.

Due to the issues stated above, the need for postprocessing the raw forecasts in the hope of improvements has gained importance in the scientific literature. A plethora of methods for statistical postprocessing exist for a range of temporal scales. These methods consist on transfer functions, computed on the basis of reforecasts, or past records of forecast-observation pairs (Hamill et al., 2004) whose goal is to match forecast values with observed ones. The choice of postprocessing method is determined by the availability of reforecast data and the application at hand. Although in principle any method could be used for seasonal forecasts, this temporal scale represents a special difficulty due to the fact that initial condition skill is mostly gone and there is little detectable signal behind a large amount of chaotic error.

In particular, for the postprocessing of ECMWF System 4 seasonal forecasts, a number of studies have been carried out: Crochemore et al., (2016); Peng et al., (2014); Trambauer et al., (2015) and Wetterhall et al., (2015). The most used methods are linear scaling and quantile mapping, although Peng et al. (2014) used a Bayesian Merging technique. In general, the aforementioned studies are successful in improving the values of the forecast qualities that they considered important. For example, Wetterhall et al., (2015) reported higher skill of forecasts of the frequency and duration of dry spells once an empirical quantile mapping has been applied to daily values of precipitation. Crochemore et al., (2016) analyzed the effect different implementations of the linear scaling and quantile mapping methods had on streamflow forecasting, concluding that the empirical quantile mapping improves the statistical consistency of the precipitation forecasts for different catchments throughout France.

The aforementioned studies have been made only for precipitation and/or mainly large areas. For hydrological applications seasonal forecasting skill of instantaneous values of precipitation (P), temperature (T) and reference evapotranspiration (E_{T0}) at the catchment scale (100 - 1,000 km^2) are, however, more important. Therefore, we analyze the bias, skill and statistical consistency of the ECMWF System 4 for Denmark focusing on P, T and E_{T0} of relevance for seasonal streamflow forecasting at catchment scale. We make use of the two most used methods for postprocessing, namely linear scaling and quantile mapping (Zhao, et al. 2017), applied to daily values. We focus on the skill of monthly aggregated values of gridded data throughout Denmark for both the raw and the corrected forecasts. We attempt to answer the following questions:

(1) what is the longest lead time for which an ‘acceptable’ forecast is achieved?

(2) is it possible to extend the acceptable forecast lead time with different postprocessing techniques?

In this study we argue that, an acceptable forecast needs to have consistency between the observed probability distribution and the predictive one, this is what we call statistical consistency throughout the paper. A statistical consistent forecast system has low (or non-existent) bias in both mean and variance. Secondly, we argue that the forecast to be used has to be better than climatology, having a higher skill both in terms of accuracy and sharpness, giving priority to the former. These characteristics for an 'acceptable forecast' follow the principle that the purpose of postprocessing is to maximize sharpness subject to statistical consistency as discussed by Gneiting, et al., (2007).

2 Data and Methods

2.1 Ensemble Prediction System and Observational Grid

Seasonal reforecast of the ECMWF System 4 for the years 1990-2013 are used in the present study. The system is comprised by 15 members (for January, March, April, June, July, September, October, December) and 51 members (for February, May, August, November) with a spatial resolution of 0.7 degrees and are run for seven months with daily output. The increase on ensemble size for February, May, August and November attempts to aid in improving forecasts for the seasons with a higher predictability. P, T and E_{T0} are the variables under study. For the computation of E_{T0}, we make use of the Makkink equation (Hendriks, 2010) that takes as inputs temperature and incoming short-wave solar radiation from ECMWF System 4.

Observed daily values for P, T and E_{T0} from the Danish Meteorological Institute (DMI) are used (Scharling and Kern-Hansen, 2012). The spatial scale for P and T, E_{T0} is 10 km and 20 km, respectively. However, we assume T and E_{T0} to be equally distributed within the 20 km and set the same values of the 20 km to the 10 km grid. Then, in total there are 662 (for P) and 724 (for T and E_{T0}) grid points that cover the 43,000 km^2 area of Denmark. Moreover, P is corrected for under catch errors as explained in Stisen, et. al., (2011) and (2012). The time and spatial variations of the variables can be seen in Fig. 1. Values are monthly accumulations for P and E_{T0} and monthly averages for T, averaged over the observed record (1990-2013). Danish weather is mainly driven by its proximity to the sea. There is a modest spatial P gradient from west to east which is more pronounced during autumn and winter. The driest month in terms of P is April and the wettest is October. E_{T0} also shows a modest spatial variability during spring and summer with larger values in eastern Denmark.

2.2 Postprocessing strategy

Given the fact that ~~both~~ the ensemble and observed spatial resolutions differ, first the ensemble forecasts were interpolated to match the 10 km grid of observed values using an Inverse Distance Weighting (Shepard, 1968), where the values at a given point of the higher resolution grid (10 km) are computed using a weighted average of the four surrounding nodes of the lower resolution forecast grid (70 km). The weights are computed as the inverse of the Euclidean distances between the observed grid node and the forecast nodes. Forecasts are then postprocessed for each grid point, time of forecast (month), and lead time (month) for each variable separately. Moreover, the computation is done in a leave-one-out cross-validation mode (Wilks, 2011 and Mason and Baddour, 2008) such that the year that we are correcting is withdrawn from the ~~training~~ sample to ensure independence between training and validation data. Then, for example for precipitation, 662x12x7x24 (# of grid points, # of months, # of lead times, # of years in the sample, respectively) correction models are computed.

Formatted: Font:

Formatted: Font:

2.3 Postprocessing methods

2.3.1 Delta method - Linear Scaling (LS)

The linear scaling approach operates under the assumption that forecast values and observations will agree in their monthly mean once a scale or shift factor has been applied (Teutschbein and Seibert, 2012). LS is the simplest possible postprocessing method as it only corrects for biases in the mean. The factor is commonly computed differently for P, E_{T0} and T due to the different nature of the variables, as P and ET0 cannot be negative.

For P and E_{T0}:

$$f^*_{k,i} = \frac{\sum_{j=1}^{N-1} y_j}{\sum_{j=1}^{N-1} \bar{f}_j} f_{k,i} \quad \text{for } i \neq j \tag{1}$$

For temperature:

$$f^*_{k,i} = f_{k,i} - \frac{1}{N-1}\left[\sum_{j=1}^{N-1} \bar{f}_j - \sum_{j=1}^{N-1} y_j\right] \qquad \text{for} \qquad i \neq j \tag{2}$$

where $f_{k,i}$ denotes ensemble member k for $k = 1,\ldots,M$ of forecast-observation pair $i = 1,\ldots,N$, M denotes the number of members (15 or 51) and *N* is the number of forecast-observation pairs, $\bar{f}_j$ denotes the ensemble mean, y_j denotes the verifying observation. Note that, as stated in Sect. 2.3., both the means of $\bar{f}_j$ and y_j are computed with the sample that withdraws forecast and observation pair *i*. Finally, $f^*_{k,i}$ represents the corrected ensemble member. Note that for precipitation, before applying the correction factor, we set all values of daily precipitation below a specific threshold to zero in order to remove the 'drizzle effect' (Wetterhall, et al., 2015). The threshold was chosen so that the number of dry days on a given forecast month matches the number of observed dry days. This threshold varies spatially with an average value of 1.5 mm/day over Denmark.

2.3.2 Quantile mapping (QM)

QM relies on the idea of Panofsky and Brier (1968). This method matches the quantiles of the predictive and observed distribution functions in the following way:

$$f^*_{k,i} = G_i^{-1}\left(F_i\left(f_{k,i}\right)\right) \tag{3}$$

where F_i represents the predictive cumulative distribution function (CDF) for forecast-observation pair *i*, G_i represents the observed CDF. Again, note that, as stated in Sect. 2.3., both F_i and G_i are computed with a sample that withdraws forecast and observation pair *i*.

F_i is calculated as an empirical distribution function ~~trained~~ fitted with all ensemble members of daily values of a given month for a given lead time and grid point. For example, for a forecast of target month June initialized in May, F_i is ~~trained~~ fitted using a sample comprising 30 (days) times 23 (number of years in the reforecast minus the year to be corrected) times 51 (number of ensemble members). The same is done for G_i, except that the ~~training~~ fitting sample is comprised by 30 x 23 values only. *F* and *G* are computed as an empirical CDF. Linear interpolation is needed in order to approximate the values between the bins of *F* and *G*. Extrapolation is then needed to map ensemble values and percentiles that are outside the ~~training~~ fitting range. Note that other approaches to deal with values outside the sample range exist that are more suitable when the focus of the study is the extreme values. For example, Wood, et al. (2002) fitted an extreme value distribution to

Formatted: Font:

Formatted: Font: Danish

Field Code Changed

Formatted: Font:

Formatted: Font:

Field Code Changed

Formatted: Font:

Formatted: Font:

Formatted: Font: English (United States)

Formatted: English (United States)

Formatted: Font:

Field Code Changed

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

extend the empirical distributions of the variables of interest. However, ~~the study of~~analyzing the effects ~~o~~if different fitting strategies is out of the scope of the present paper.

2.4 Verification metrics

As a manner to evaluate first the raw forecasts and the improvement after postprocessing we check for four qualities: bias, skill in regards to accuracy and sharpness, and statistical consistency.

2.4.1 Bias

Bias is a measure of under – overestimation of the mean of the ensemble in comparison with to the observed mean:

$$\%Bias = \left(\frac{\sum_{i=1}^{N} \bar{f}_i}{\sum_{i=1}^{N} y_i} - 1 \right) \times 100 \qquad (4)$$

for P and E_{T0} and:

$$Bias = \frac{1}{N} \left[\sum_{i=1}^{N} \bar{f}_i - \sum_{i=1}^{N} y_i \right] \qquad (5)$$

for T. f_i and y_i are the same as in Eq. (1).

If the bias is negative, the forecasting system then exhibits a systematic underprediction. Conversely, if the amount is positive the system shows an average overprediction. Values closer to 0 are of course desirable.

2.4.2 Skill

The skill of a forecasting system is the improvement, on average, that the system has with respect to a reference system that could be used instead, for example climatology for seasonal forecasts or persistence for short-range forecasts. The skill score is computed in the following manner

$$Skill = \frac{Score_{sys} - Score_{ref}}{Score_{per} - Score_{ref}} \qquad (6)$$

where $Score_{sys}$, $Score_{ref}$ and $Score_{per}$ are the score value of the system to be evaluated, the reference system and the value of a perfect system, respectively. The range of the skill is from $-\infty$to 1 and values closer to 1 are preferred. In this paper we calculate the skill with respect to accuracy and sharpness. We compute the continuous rank probability score (CRPS) (Hersbach, 2000), as a general measure of the accuracy of the forecast as it contains information of both forecast biases in the mean and spread. The computation of the score is as follows:

$$CRPS = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \left[P_i(x) - H(x - y_i) \right]^2 dx \,, \qquad (7)$$

where $P_i(x)$ is the CDF of the ensemble forecast for pair *i* and $H(x-y_i)$the Heaviside function that takes the value 1 when $x > y_i$ and 0 otherwise, y_i and *N* are, as in Eq. (1), the verifying observation for forecast-observation pair *i*, and the number of forecast-observation pairs, respectively. We made use of the EnsCrps function of the R package SpecsVerification (Siegert, 2015) developed in R version 0.4-1. For the skill with respect to sharpness we use the average along $i = 1, \ldots, N$ of the differences between the 25% and the 75% percentiles of each of the ensemble CDFs, P_i.

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

Formatted: Font:

In Eq. (6) our reference forecast is ensemble climatology (1990-2013) where the year to be evaluated is withdrawn from the sample. Both the accuracy and sharpness score for a perfect system cf. Eq. (6), $Score_{per}$ is equal to zero so the skill score can be then simplified as:

$$Skill = 1 - \frac{Score_{sys}}{Score_{ref}} \tag{8}$$

which, once multiplied by 100, can be seen as the percentage of improvement (if positive) or worsening (if negative) over the reference forecast. Throughout the paper, the skill related to accuracy will be denoted as CRPSS whereas the skill due to sharpness will be denoted as SS.

Furthermore, in order to define the statistical significance of the differences between the skill of ensemble climatology and ECMWF System 4 forecasts, as well as the postprocessed predictions, a Wilcoxon-Mann-Whitney test (WMW-test; see Hollander et al, 2014) was carried out. The WMW test, unlike the most common t-test, makes no assumptions about the underlying distributions of the samples. We applied the test for each grid point, target month and lead time.

2.4.3 Statistical consistency

We use the Probability Integral Transform (PIT) diagram for a depiction of the statistical consistency of the system. The PIT diagram is the CDF of the z_i's defined as $z_i = P_i(X \le y_i)$. Therefore, z_i is the value that the verifying observation y_i attains within the ensemble CDF, P_i. The diagram represents an easy check of the biases in the mean and dispersion of the forecasting system. For a forecasting system to be consistent, meaning that the observations can be seen as a draw of the forecast CDF~~, a quality that a forecasting system should aim for~~, the CDF of z_i should be close to the CDF of a uniform distribution on the [0,1] range. Deviations from the 1:1 diagonal represent bias issues in the ensemble mean and spread. The reader is referred to Laio and Tamea (2007) and Thyer, et al., (2009) for an interpretation of the diagram. ~~as explained in appendix A and shown in Fig. A12~~. Similar to (Laio and Tamea, 2007), we make use of the Kolmogorov bands to have a proper graphical statistical test for uniformity. Finally, we make use of the Anderson-Darling test (Anderson and Darling, 1952) for a numerical test of the uniformity of the PIT diagrams. We carry out the test using the ADGofTest (Gil, 2011) R package (R Core 2017). Here, the null hypothesis is that the PIT diagram follows a uniform distribution on the [0-1] range.

2.5 Accuracy of maximum monthly daily precipitation and number of dry days

For applications such as flooding and forecasting of low flows and droughts, water managers might be interested not only in the skill of monthly accumulated precipitation but also in the skill of other precipitation quantities. We will use a rather simple approach ~~for~~ to ~~checking the~~check for deficiencies of the raw forecasts and whether the postprocessing methods improve these deficiencies. We will ~~by analyzing~~ analyze the improvement in the prediction of ~~also~~ monthly maximum daily precipitation and number of dry days in a given month. For the purpose of this study, a dry day is defined as the day with observed zero-precipitation, and the comparison with the ensembles is made on a daily basis.

3 Results

3.1 Analysis of raw forecasts

The first row in Fig. 2 depicts the ECMWF System 4 forecast and ~~an~~ the ensemble climatological forecast for August accumulated P and E_{T0} and averaged T for one grid located in west-central Denmark for the first month lead time. The values for different forecast qualities for that grid point are also included. For a forecasting system to be useful, it has to be at least, better than a climatological forecast. For the given example here, ~~we show in the background~~ the reference forecast ~~that~~ is wider than the ECMWF System 4 forecast. This is an example of a month where we have a slightly better skill than the ensemble climatological forecast for the three variables in question. For example, raw T predictions from ECMWF System 4 improve, on average, on the reference forecast by 22~~0~~% in terms of accuracy. This level of skill is attained due to the sharper forecasts that exhibit a low bias (-0.23 deg C). On the other hand, sharpness is only a desirable property when biases are low. This is illustrated for P forecasts attaining a high skill due to sharpness (0.43) but at the expense of a low skill due to accuracy (0.01) that is caused by the high negative bias (-14.12%) where, for example for 1992 and 2010, the verifying observation lies outside the ensemble range contributing negatively to the CRPS in Eq. (7).

3.1.1 Bias

In an effort to summarize the results, a spatial average of the bias throughout Denmark was computed. Figure 3 shows the spatial average bias of P, T and E_{T0} of the raw forecasts. Y-axis represents the target month, for example April, and X-axis represents the forecast lead time, lead time 5 is the forecast for April initiated in December. As we can see from Fig. 3, bias depends on the target month and, to a lesser degree, on lead time. For P, the lowest bias can be found throughout autumn to beginning of winter, followed by a general underestimation of P that is at its highest for June. April shows an overestimation that might be due to the 'drizzle effect' in a month where dry days are slightly more common, in comparison to March and May (the percentage of dry days within the month is 50%, 57% and 56% in March, April and May, respectively). The drizzle effect issue is a very well-known problem of GCM, and is related to the generation of small precipitation amounts, usually around 1.0-1.5 mm/day where observed precipitation is not present (Wetterhall et al., 2015).

T bias averaged over Denmark has a range that lies within [-2,2] degrees Celsius. The bias switches from positive to negative when temperatures start to increase in March and from negative to positive bias when temperatures start decreasing in August. This indicates that the forecast of T has a smaller annual amplitude than observed. Lowest biases are encountered during January and February with a bias of 0.5 deg Celsius, and it is higher during late spring and summer, with a negative bias of almost 2 deg Celsius. Finally, the bias range for E_{T0} is smaller than for P taking values within [0-25%] on average over Denmark. In general, there is a positive bias, which is at its highest during February.

However, averaging does not tell the whole story. We are also interested in the spatial variation of biases over Denmark. Figure 4 shows the spatial distribution of bias for the first month lead time and its evolution during summer. In general, there is an underestimation of P throughout Denmark, much more pronounced during June. Nevertheless there also exists a positive bias in central Jutland and on the urban area of Copenhagen reaching a value around 10-20%. The positive bias area grows in July occupying most of Jutland and North Zealand.

Other seasons were also mapped and shown in the supplements as Fig. S1 to Fig. S3. During autumn and winter there is also a general negative bias that is more pronounced in central Jutland, reaching values of -30%. Nevertheless, an overestimation exists in eastern Denmark for those seasons. For winter, this overestimation is present in the sea grid points. Finally, during spring the spatial variability changes. For example, most grid points exhibit a positive bias during April, except for the southeast region of Denmark that has a small negative bias between 0.0 to 5.0%. During May a tendency of overforecasting is present in central Jutland.

The spatial distribution of T bias during autumn and winter (Fig. S2 and Fig. S3, respectively) follow a similar pattern with a general positive bias reaching its highest values in the southeast region (from 1.5 to 2.0 deg C). A negative bias is seen during spring (Fig. S1) and late summer across Denmark (Fig. 4). In June a positive bias [0-2 deg C] is present in a large area of the Jutland peninsula (Fig. 4). Finally, the spatial variation of bias of E_{T0} is less pronounced and, in general, positive. Nonetheless, exceptions exist. There is a negative bias in small regions located in the coastal areas or sea grid points, that ranges from -10 to 0%.

The results presented above are specifically for lead time 1, i.e., forecasts of accumulated P and E_{T0} and average T for August initialized on August 1st. The spatial variation of bias for other lead times was also mapped (not shown) and analyzed. In general, similar spatial patterns were found for all three variables, being the same along the target months regardless of lead time.

3.1.2 Skill

Figure 5 summarizes the results for skill in the following manner. First, the values presented are, like in Fig. 3, the spatial average of skill for entire Denmark. Secondly, we compare the skill in terms of accuracy (first row) and sharpness (second row) ~~and visualize it in a format that can give us both skill comparisons at the same time. Accuracy and sharpness skill scores were binned into four categories with the first one being the case where, on average, ECMWF System 4 is scoring worse than a forecast based on ensemble climatology. The remaining bins represent the different levels of skill where System 4 scores better than the reference forecast.~~

~~The bins for skill with respect to accuracy are represented with different colors, whereas the bins for skill with respect to sharpness are represented with different shapes/sizes, the bigger the size the wider the ensemble from System 4 is with respect to ensemble climatology, on average. The squares represent the case where ECMWF System 4 has a negative skill score, indicating that climatological forecasts are sharper than those of System 4~~. As for bias, skill appears to be dependent on the target month and, to a lesser degree, on lead time.

For P, and looking at the first month lead time, ECMWF System 4 skill in accuracy mildly beats that of climatology for February, March, April, July, August, November and December, with a CRPSS of 0.15 at most. In general, skill in accuracy decreases for lead time 2 onwards. April stands out for having a slightly positive skill in accuracy for almost all lead times, but comes at the expense of having a wider spread than ensemble climatology. For T, for the first month lead time, a positive skill exists in terms of accuracy for almost all months, except for late spring. Skill in terms of accuracy also decreases with lead time, but February stands out to have a mild skill for almost all lead times. Forecasts are also in general sharper than ensemble climatology, with the exception of January and March at longer lead times. E_{T0} appears to have skill only for late summer and beginning of autumn in terms of accuracy. This may be explained by the fact that forecasts are sharper than climatology, indicating that there could be an underdispersion issue.

Note that the computation of the skill score for accuracy was done using a $CRPS_{sys}$ of 15 or 51 ensemble members, while the $CRPS_{ref}$ is comprised by 23 members. The disparity in the number of ensemble members can cause forecasts to be in disadvantage (advantage) compared to a reference forecast with larger (smaller) ensemble size. Ferro, et al., (2008) provided estimates of unbiased skill scores that take into account the differences in ensemble size. In an attempt to remove this effect, we calculated the CRPSS using the unbiased estimator for $CRPS_{ref}$ in Eq. (22) in Ferro, et al., (2008). In general, there was a mild increase in the CRPSS value for the months with 15 members, as expected (not shown). The opposite holds for the months with 51 members (February, May, August and November), ~~there was~~where a mild decrease of the CRPSS values (not shown) was obtained. ~~Due to the fact that~~Because the changes in CRPSS are moderate, in the rest of the document we will report the results of the CRPSS using the original ensemble sizes.

Formatted: Subscript

Figure 6 shows the skill in accuracy (CRPSS) for monthly values and for lead time of one month mapped across Denmark and its monthly evolution during the summer. The rest of the seasons are also mapped and analyzed. These are included in the supplement as Fig. S4 to Fig. S6. Higher skill is observed for T, for which ECMWF System 4 improves the ensemble climatological forecast with up to 50%. P and E_{T0} have lower skill in comparison to T, reaching a value of 0.3 for specific months and regions in Denmark. The spatial variation of skill for P seems scattered across Denmark and also through the year. Some notable exceptions are the higher skill in accuracy that ECMWF System 4 has in western Jutland during November and December and the low skill attained in October across Denmark (Fig. S5 and Fig. S6). The spatial variation of skill in accuracy of monthly averaged T seems more pronounced during autumn and spring, with northern Denmark attaining the highest values of skill for these seasons. For the remaining seasons, skill over climatology is present across the country, except for late spring where eastern Denmark has the largest negative skill. Finally, the spatial variation of skill of E_{T0} is more pronounced for the months April to November with both positive and negative skill. In general, in this period eastern Denmark attains positive, although mild for some months, values of skill, except for November.

In general, the areas with highest biases shown in Fig. 4 are associated with the lowest skill scores. For instance, for October P in southern central Jutland the negative bias reaches values around 30-40 % (Fig. S2), leading to values of CRPS almost 60% smaller than that of ensemble climatology (Fig. S5). The opposite also holds, areas where biases are lower, tend to have the highest benefits over ensemble climatology, i.e., March P across Denmark or November P in western Jutland.

Skill related to accuracy was also mapped for lead times 2-7 months (not shown). In general, regions having a statistically significant positive skill score for lead time 1 month vanish, except for some smaller regions where a slight positive skill, between 0.0 and 0.1 is found, i.e. April P forecast initiated in February (lead time 3 months), which contributes to the mild positive skill at longer lead times as seen in Fig 5.

The skill related to sharpness was also mapped for all target months and lead times (not shown). In general, forecasts are sharper than ensemble climatology as seen in Fig. 5 across Denmark for all three variables under study. This situation persists, in general, along all lead times, except for April and October P and January, March and November T, as shown also in Fig. 5. For these months, the lack of sharpness is present th~~o~~rough Denmark. Nevertheless, for April P, the region with the lack of sharpness is located in southern Denmark, along all lead times.

3.1.3 Statistical consistency of monthly aggregated P, T and E_{T0}

~~We follow Fig. A12 in Appendix A for the interpretation of the PIT diagram.~~ The first row in Fig. 7 shows the PIT diagram for raw ECMWF System 4 summer forecasts for lead time 1 month. The observations and forecast of the diagram come from a grid point located in western Denmark (squared shape in Fig. 8). The remaining seasons can be seen in Fig. S7 to Fig. S9 in the supplementary material. Raw P forecasts, for this particular grid point, exhibit an underprediction of the mean for winter~~,~~ ~~summer~~ and autumn, and with mixed results for the remaining seasons. For example, ~~This~~ the underprediction bias is somehow reduced for spring, except for April, where the system exhibits a positive bias. Raw T predictions of winter, October and November ~~and autumn~~, in addition to June, exhibit an overprediction, which is lowest for January and February. Spring and summer T exhibit an underprediction which is highest for ~~April~~ July (Fig. S7 and Fig. 7).

Finally, raw forecasts of E_{T0} during all seasons, exhibit an ~~underdispersion~~overprediction of the mean, in accordance with the results in Sect. (3.1.1~~,~~). The statistical consistency at longer lead times for all variables (2-7 months, not shown), depends, similarly to bias, on the target month and, to a much lesser degree on~~regardless of~~ lead time.

~~i.e. the majority of the verifying observations lie on the tails or outside the ensemble range. This is a consequence of a forecast with insufficient spread. Note, however that the underdispersivity is not present in selected months: June, November and February.~~

Issues with bias in the ensemble spread ~~of P and T are not remarkably~~are only clear for selected months of P and T. For example, March T and, to a lesser degree, July P exhibit underdispersion issues, i.e. too often, the observations lie outside the ensemble range. ~~clear from the visualization of the PIT diagrams in Fig. 7. This situation is perhaps explained by the fact that we are analyzing the statistical consistency of monthly aggregated values which smoothes out extremes.~~ ~~The statistical consistency at longer lead times for all variables (2-7 months, not shown), depends, similarly to bias, on the target month regardless of lead time.~~

Note that the analysis of statistical consistency is done for one grid point. In Sect. 3.2.3. we will aggregate the results to include the uniformity test across Denmark.

3.2 Analysis of postprocessed forecasts

The second and third ~~columns column~~rows in Fig. 2 show the corrected forecasts and the bias and skill scores after postprocessing using the LS and the QM method, respectively. The results represent a particular grid point and forecast of August initialized August 1. After postprocessing, the reduction of bias is evident for the three variables under study. Nevertheless, and contrary to what one should expect, this reduction of bias does not necessarily translate into an increase of skill in accuracy, at least for P and T and for this particular month and grid point. The quantification of the reduction/increase of accuracy after postprocessing for the whole Denmark, through the year and for different lead times is discussed in the Sect. 3.2.2 below.

3.2.1 Bias

Any postprocessing technique used should be able to at least remove biases in the mean. This is accomplished using both techniques. Figure 8 shows the bias of P, T and E_{T0} and its evolution through the year for lead time 1 month. Bias is shown for four locations scattered around Denmark. Figure 8 shows that the yearly variability of the bias is collapsed to almost 0%, although for P and winter E_{T0}, the LS method seems to be doing a slightly better job at removing the bias than the QM. This comes as no surprise as the LS method forces this bias to be zero.

3.2.2 Skill

In order to be more quantitative in terms of the improvement over the raw forecast, we counted the number of grid points for which the skill score was positive and the number of grid cells for which the skill score was negative. Furthermore, the scores are only considered positive or negative if the differences in the distribution of the skill between ensemble climatology and ECMWF System 4 forecasts are statistical significant at the 0.05 level using the WMW test. Consequently, we introduced a third category for which there is no statistical significant difference in skill between climatology and ECMWF System 4 forecasts.

Figure 9 shows the percentage of grid cells with a statistical significant positive skill due to accuracy, Eq. (8), for the raw forecasts (first raw) and the postprocessed forecasts (LS, second row; QM third row). All target months and lead times are included. If a postprocessing method is successful in increasing the regions with positive skill scores, then the box for that particular target month/lead time is bluer in comparison to the raw forecasts. For P, there is no obvious increase in skill due to accuracy, except, perhaps, February and July forecasts for the first month lead time. There are, however, instances for which the percentage of positive skill grid points decreases. The most obvious cases are March and November (1st month lead time) with a reduction of almost half, i.e., from 13.6% (raw) to 5.7% (LS) for March. On the contrary, T and E_{T0} exhibit a greater improvement, at least on the first month lead time. For instance, the percentage of grid points with positive skill increases from 4.5% to 50% for April T (LS) and from 30% to 100% (LS) for July T. The biggest improvement for E_{T0} appears in June (first month lead time), reaching 90% of positive grid cells after postprocessing (LS).

In addition, the negative and equal categories were also plotted and included in the supplement as Fig. S10 and Fig. S11. After postprocessing, there are instances where a considerable amount of grid cells move from a statistical significant negative skill score to the third category (no significant differences between ensemble climatology and ECMWF System 4 score distributions, Fig. S11). This is true for T and E_{T0} at longer lead times. One of the obvious examples is February E_{T0} at lead time 6 (forecast initiated in September), the percentage of grid points with negative skill scores decrease from 80.5% to 4.3% after postprocessing (Fig. S10). On the other hand the percentage of grid points with no significant differences in skill increase from 20% to 95.7% after postprocessing (Fig. S11) for this particular example.

To further illustrate the above situation, Fig. 10 shows the spatial distribution of skill due to accuracy encapsulated in box-plots that represent the 662/724 grid points across Denmark. Figure 10 shows the CRPSS for the target month of February at all lead times and the raw and postprocessed skill. The figure shows a reduction of the spatial variability of skill in accuracy and for this particular month, this reduction is more pronounced for E_{T0}. However, and as mentioned above, the reduction of spatial variability of accuracy is not enough to ensure statistical significant positive differences in skill.

We also constructed ~~Fig. 9~~figures~~ but~~ for sharpness (Fig. S12 to Fig. S14) similar to Fig. 9. It is evident that a loss of sharpness occurs after postprocessing in comparison to the raw forecasts for LS and QM applied to P, and QM applied to T and E_{T0}. Sharpness seems to be maintained for T and E_{T0} when we use the LS method. This can be explained by the fact that the correction factor applied to T forecasts is additive, which in turn changes the level of the ensemble members and has no effect in the spread of the forecasts, leaving the sharpness score equal to that of the raw forecasts. On the other hand, when the correction factor is multiplicative, as in Eq. (1) for P and E_{T0}, not only the level but the spread is affected. It will increase the spread when the correction factor is above 1 (which indicates an underprediction issue), and conversely, reduce the spread when the correction factor is below 1 (indicating an overprediction issue). The larger the correction factor is the larger effect it will have in the ensemble spread. This explains why for E_{T0}, where biases are in general lower than biases in P, sharpness seems not to be affected. This effect is somewhat artificial and may lead to misleading evaluations of the power LS has in correcting for biases in spread.

3.2.3 Statistical consistency of postprocessed monthly aggregated forecasts

Second and third row in Fig. 7 and Fig. S7 to S9 show the PIT diagrams of corrected forecasts for one grid point located in western Denmark. In general, the statistical consistency seems to be improved (points closer to the 1:1 diagonal in Figure 7) to the same degree for both postprocessing methods. Although, for E_{T0}, this consistency is better enhanced by QM. This fact may be explained by the more evident sharpness ~~lost~~ loss after correcting forecasts with the QM method~~of sharpness that QM has~~ (Fig. S11 to S13). ~~in an attempt of adjusting the biases in spread present in the raw forecasts (Fig. 7).~~

Fig. 11. depicts the results of the AD test in the following manner. First, the first, second and third rows represent the results of the raw, and postprocessed forecasts with LS and QM methods, respectively. Secondly, as for Fig. 9, the x-axis represents the lead time and the y-axis the target month. Finally, the percentage shows the proportion of grid points for which the null of uniformity at the 5% significance level is accepted. A variety of results are found by the inspection of Fig. 11. First, the percentage of grid points for which the uniformity hypothesis is accepted is very low for raw forecasts. This conclusion holds except for T in January and February with forecasts initialized in months with 51 ensemble members. Secondly, the percentage increases after postprocessing for selected months and lead times, usually involving target months with 51 ensemble members. This increase is more visible in postprocessed forecasts using the QM method. Finally, the statistical consistency of E_{T0} appears to remain low even after postprocessing.

Formatted: Subscript

3.2.4 Accuracy of extreme precipitation and number of dry days

Figure 12~~1~~ shows the skill in terms of accuracy for both monthly maximum precipitation and the monthly number of dry days. Box-plots represent the distribution of the skill score of all 662 grid cells. Skill scores are for January forecasts for all seven lead times. Two features are highlighted, first, spatial variability of skill gets reduced after postprocessing and secondly, for the skill of the number of dry days, results show that QM performs significantly better than LS. This is not surprising as QM adjusts for biases in the whole range of percentiles of the distributions, whereas LS only focuses on the mean. Note that the apparent increase in skill after LS postprocessing is a consequence of the drizzle effect removal before bias correction. Despite the reduction of the spatial variability and an increase, on average, of the skill of postprocessed monthly maximum P and number of dry days, results still show a difficulty to beat climatology, as the CRPSS is still negative, even after bias corrections are implemented.

4 Summary and Conclusions

The present study had two objectives. The first one was to analyze the bias and skill of the ECMWF System 4 in comparison to a climatological ensemble forecast, i.e. a forecast based on observed climatology over a period of 24 years, and well as comparing the statistical consistency between the predictive distribution and the distribution of ~~the~~ its verifying observations. This analysis was done for hydrological relevant variables P, T and E_{T0}. The conclusions of the first objective of the study and that answer the first question posed in ~~the Sect.~~section 1 can be summarized as follows:

- Raw seasonal forecasts of P, T and E_{T0} from ECMWF System 4 exhibit biases that depend on the target month and to a lesser extent, on lead time. This result is also in accordance to what was found in Crochemore et al., (2016). There is a persistent overforecasting issue for E_{T0}, which combines biases of both T and incoming shortwave solar radiation.
- ~~In addition to the biases, Crochemore et al., (2016) also found a rather similar degree of skill of the raw ECMWF System 4 forecasts for mean areal P in France.~~ In general, skill in terms of accuracy is only present during the first month lead time, which is basically the skill of the medium-range forecast. Crochemore et al., (2016) found a similar degree of skill of the raw ECMWF System 4 forecasts for mean areal P in France.
-

One seeming advantage ECMWF System 4 has over ensemble climatology is that forecasts are sharper. However, ~~This~~ this overconfidence, combined with the biases in the mean lead to lower levels of accuracy in comparison to the accuracy of the ensemble climatological forecasts.

- Using the PIT diagrams we were able to confirm the results for the bias on the mean of P~~, and~~ T and E_{T0}.

~~Bias in spread are present for E_{T0}, particularly for the months with the lowest number of ensemble members (15).~~

The second objective was to improve the forecasts using two relatively simple methods of postprocessing: LS and QM. This was done having in mind the problems GCMs have with regards to both bias in the mean and dispersion. Modest improvements were found and can be summarized as follows:

- Both methods ~~act equally good~~perform equally well in removing biases in the mean.
- In terms of accuracy, mild improvements are seen on the first month lead time, especially for T and E_{T0}, where a higher portion of grid points are able to reach a positive skill. P and longer lead times are still difficult to improve. This may be explained by the same situation as discussed in Zhao, et al., (2017). QM ~~assumes~~ performs better when~~that~~ there ~~is~~ exist a linear relationship between ensemble mean and observations. This linear relationship may be absent~~, assumption may not hold~~ at longer lead times reducing the effectiveness of these methods.
- Looking at the spatial distribution of skill in sharpness we see that for P, both methods tend to decrease it, with a slight increase of QM over LS. For T and E_{T0}, LS seems to be able to keep the sharpness of the raw forecasts. This is not the case for QM, for which for some months it manages to disappear the areas where a slight positive skill is present. Note, however, that sharpness using the LS method ~~gets~~ is improved when the correction factor is multiplicative and less than one (positive bias; i.e., E_{T0}). The opposite holds, sharpness is inflated when the multiplicative correction factor is larger than one (negative bias; i.e., P). This has implications for the computation of the CRPS, as it also penalizes (rewards) for wide (narrow) ensemble forecasts, on top of the penalization for biased predictions. This situation may also explain why in Crochemore, et. al., 2016, LS has a better improvement in terms of sharpness than QM, at least for spring P.

Formatted: Not Highlight

Formatted: Don't add space between paragraphs of the same style, Bulleted + Level: 1 + Aligned at: 0,63 cm + Indent at: 1,27 cm

Formatted: Font: (Default) Times New Roman, 10 pt

Formatted: Subscript

Formatted: Font: (Default) Times New Roman, 10 pt

Formatted: Add space between paragraphs of the same style, No bullets or numbering

- Statistical consistency is ~~better~~ improved ~~for~~ using QM~~ and for E_{T0} forecasts that exhibit biases in the ensemble spread~~. Moreover, QM also performs better in correcting ~~for~~ biases of low values of P. This is not a surprising result, as QM corrects for biases for the entire percentile range.

We are aware that our research may have limitations. The first is that methods applied here were implemented on a grid-to-grid basis that may not correct for displacements and might lose spatio-temporal and intervariable dependencies. Spatial correction methods have been suggested such as the ones used by Feddersen and Andersen, (2005) and Di Giuseppe et al., (2013). Another suggestion has been to recover these dependencies by adding a final postprocessing step such as the methods proposed in Clark et al., (2004) or Schefzik et al., (2013). The second is ~~that~~ the exclusion of postprocessing methods tailored to ensemble forecasts that take into account the joint distribution of forecasts and observations (Raftery, et. al., 2005; Zhao et al., 2017). Their inclusion would gain a deeper insight to the comparison presented here by increasing the complexity of the correction methods and the evaluation of their added value in comparison to simpler approaches.

Postprocessing for seasonal forecasting is still a subject at its infancy, and although one could argue that advances in seasonal forecasting will make postprocessing unnecessary in the future, there is still a long way to go to get there. GCMs suffer from several issues as discussed here, however, we still encourage its use. They are physically-based, sharper than climatological forecasts~~ with ensemble climatology,~~. ~~and~~ We believe that once biases issues~~ with their biases discussed here~~ are fixed by means of a more realistic representation of coupled and subgrid processes and/or a better integration of observational data using~~ an updated~~ data assimilation~~ procedure~~ (Weisheimer and Palmer, 2014; Doblas-Reyes et al., 2013), they will be able to provide valuable information at longer lead times for sector applications such as water management.

~~5 Appendix A. Probability Integral Transform Diagram~~

~~The interpretation of the shape of the PIT diagram is based on Fig. A12 modified from (Laio and Tamea, 2007). Deviations from the 1:1 diagonal, point to the different biases in the mean and dispersion. Four situations can arise:~~

~~1. Overprediction, or positive bias in the mean: The CDF of the *zi's* lies above the 1:1 diagonal.~~

~~2. Underprediction, or negative bias in the mean: The CDF of the *zi's* lies below the 1:1 diagonal.~~

~~3. Overdispersion, or positive bias in spread (underconfident): A greater proportion of the values of the CDF lie on the middle ranges bins of the distribution.~~

~~4. Underdispersion, or negative bias in spread (overconfidence): A greater proportion of the values of the CDF lie on the tails of the distribution.~~

~~6~~ 5 Acknowledgements

This study was supported by the project 'HydroCast - Hydrological Forecasting and Data Assimilation', Contract #0603-00466B (http://hydrocast.dhigroup.com/) funded by the Innovation Fund Denmark. Special thanks to Florian Pappenberger for providing the ECMWF System 4 reforecast, ~~and~~ to Andy Wood and Pablo Mendoza for hosting the first author at NCAR and to Louise Crochemore (SMHI), for aiding in the implementation of the unbiased CRPSS. Finally, the authors would like to thank three anonymous reviewers for providing feedback that certainly increased the quality of the work.

~~7~~ 6 References

Anderson, T.W. and Darling, D.A.: Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics, 193–212, 1952.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R.: The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields, J. Hydrometeorol., 5(1), 243–262, doi:10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2, 2004.

Formatted: Space Before: 0 pt, After: 10 pt, Line spacing: Multiple 1,15 li, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font: Not Bold

Crochemore, L., Ramos, M.-H. and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 20, 3601-3618, doi:10.5194/hess-20-3601-2016, 2016.

Dessai, S. and Soares, M. B.: European Provision Of Regional Impact Assessment on a Seasonal-to-decadal timescale. Deliverable 12.1 Systematic literature review on the use of seasonal to decadal climate and climate impacts predictions across European sectors, Euporias, 12(3082911), 1-26, available at: http://www.euporias.eu/system/files/D12.1_Final.pdf (last access: 1 June 2017), 2013.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P. and Rodrigues, L. R. L.: Seasonal climate predictability and forecasting: Status and prospects, Wiley Interdiscip. Rev. Clim. Chang., 4(4), 245–268, doi:10.1002/wcc.217, 2013.

Feddersen, H. and Andersen, U.: A method for statistical downscaling of seasonal ensemble predictions, Tellus, Ser. A Dyn. Meteorol. Oceanogr., 57(3), 398–408, doi:10.1111/j.1600-0870.2005.00102.x, 2005.

Di Giuseppe, F., Molteni, F. and Tompkins, A. M.: A rainfall calibration methodology for impacts modelling based on spatial mapping, Q. J. R. Meteorol. Soc., 139(674), 1389–1401, doi:10.1002/qj.2019, 2013.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 19–24, doi:10.1002/met.45, 2008.

Gil, C.: ADGofTest. Anderson-Darling GoF test. R package version 0.3. https://CRAN.R-project.org/package=ADGofTest. Last access: 31-07-2018, 2011.

Gneiting T, B. F. and AE, R.: Probabilistic Forecasts, Calibration and Sharpness, J. R. Stat. Soc. Ser. B (Statistical Methodol., 69, 243-268, 2007.

Hamill, T. M., Whitaker, J. S. and Wei, X.: Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts, Bull. Am. Meteorol. Soc., 3825–3830, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2, 2004.

Hendriks, M.: Introduction to Physical Hydrology, Oxford University Press, 2010.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15(5), 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hollander, M., Wolfe, D. A. and Chicken, E.: Nonparametric statistical methods, 3rd edn., Wiley Series in Probability and Statistics, 2014.

Hudson, D., Marshall, A. G., Yin, Y., Alves, O. and Hendon, H. H.: Improving Intraseasonal Prediction with a New Ensemble Generation Strategy, Mon. Weather Rev., 141(12), 4429–4449, doi:10.1175/MWR-D-13-00059.1, 2013.

Kim, H. M., Webster, P. J., Curry, J. A. and Toma, V. E.: Asian summer monsoon prediction in ECMWF System 4 and NCEP CFSv2 retrospective seasonal forecast, Clim. Dyn., 39(12), 2957–2991, doi:10.1007/s00382-012-1470-5, 2012a.

Kim, H. M., Webster, P. J. and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, Clim. Dyn., 39(12), 2957–2973, doi:10.1007/s00382-012-1364-6, 2012b.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sc., 11, 1267–1277, 2007.

Liu, X., Wu, T., Yang, S., Jie, W., Nie, S., Li, Q., Cheng, Y. and Liang, X.: Performance of the seasonal forecasting of the

Formatted: Line spacing: Multiple 1,15 li, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Line spacing: At least 7,5 pt, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers, Pattern: Clear (White), Tab stops: 1,62 cm, Left + 3,23 cm, Left + 4,85 cm, Left + 6,46 cm, Left + 8,08 cm, Left + 9,69 cm, Left + 11,31 cm, Left + 12,92 cm, Left + 14,54 cm, Left + 16,16 cm, Left + 17,77 cm, Left + 19,39 cm, Left + 21 cm, Left + 22,62 cm, Left + 24,23 cm, Left + 25,85 cm, Left

Formatted: Font: Font color: Black, Check spelling and grammar, Border: : (No border)

Asian summer monsoon by BCC_CSM1.1(m), Adv. Atmos. Sci., 32(8), 1156–1172, doi:10.1007/s00376-015-4194-8, 2015.

Maclachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., Gordon, M., Vellinga, M., Williams, A., Comer, R. E., Camp, J., Xavier, P. and Madec, G.: Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system, Q. J. R. Meteorol. Soc., 141(689), 1072–1084, doi:10.1002/qj.2396, 2015.

Mason, S. J. and Baddour, O.: Statistical Modelling, in: Seasonal Climate: Forecasting and Managing Risk, Springer 82, 167-206, 2008.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T. and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Tech. Memo., 656, 49 pp., available at: https://www.ecmwf.int/sites/default/files/elibrary/2011/11209-new-ecmwf-seasonal-forecast-system-system-4.pdf (last access: 1 June 2017), 2011.

van Oldenborgh, G. J., Balmaseda, M. A., Ferranti, L., Stockdale, T. N. and Anderson, D. L. T.: Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-year period, J. Clim., 18(16), 3250–3269, doi:10.1175/JCLI3421.1, 2005.

Panofsky, H. W. and Brier, G.W.: Some Applications of Statistics to Meteorology, The Pennsylvania State University Press, Philadelphia, U.S., 1968.

Peng, Z., Wang, Q. J., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P. and Wang, Z.:Statistical Calibration and Bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China, J. Geophys. Res. Atmos., 119, 7116–7135, doi:10.1002/2013JD021162, 2014.

R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Last access: 31-07-2018, 2017.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133(5), 1155–1174, doi:10.1175/MWR2906.1, 2005.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M. and Becker, E.: The NCEP Climate Forecast System Version 2, J. Clim., 27(6), 2185–2208, doi:10.1175/JCLI-D-12-00823.1, 2013.

Scharling, M. and Kern-Hansen, C.: Climate Grid Denmark - Dateset for use in research and education, DMI Tech. Rep., (10), 1–12 [online] Available from: www.dmi.dk/dmi/tr12-10, 2012.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T.: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling, Stat. Sci., 28(4), 616–640, doi:10.1214/13-STS443, 2013.

Shepard, D. S.: A two dimensional interpolation function for irregularity spaced data. Proceedings of the 23rd Associations for Computing Machinery Conference, ACM, 517-524, 1968.

Siegert, S.: SpecsVerification: Forecast verification routines for the SPECS FP7 project, R package version 0.4-1, available at: https://cran.r-project.irg/web/packages/SpecsVerification/SpecsVerification.pdf (last access: 12-June-2017), 2015.

Stisen, S., Sonnenborg, T. O., Højberg, A. L., Troldborg, L. and Refsgaard, J. C.: Evaluation of Climate Input Biases and Water Balance Issues Using a Coupled Surface–Subsurface Model, Vadose Zo. J., 10(1), 37, doi:10.2136/vzj2010.0001, 2011.

Formatted: Line spacing: At least 7,5 pt, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers, Pattern: Clear (White), Tab stops: 1,62 cm, Left + 3,23 cm, Left + 4,85 cm, Left + 6,46 cm, Left + 8,08 cm, Left + 9,69 cm, Left + 11,31 cm, Left + 12,92 cm, Left + 14,54 cm, Left + 16,16 cm, Left + 17,77 cm, Left + 19,39 cm, Left + 21 cm, Left + 22,62 cm, Left + 24,23 cm, Left + 25,85 cm, Left

Formatted: Font: Font color: Black, Check spelling and grammar, Border: : (No border)

Stisen, S., Hojberg, A. L., Troldborg, L., Refsgaard, J.C., Christensen, B. S. B., Olsen, M. and Henriksen, H. J.: On the importance of appropiate precipitation gauge catch correction for hydrogical modelling at mid to high altitudes, Hydrol. Earth Syst. Sci., 16(11), 4157-4176, doi:10.5194/hess-16-4157-2012, 2012.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, J. Hydrol., 456–457, 12–29, doi:10.1016/j.jhydrol.2012.05.052, 2012.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and Srikanthan, S: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45, W00B14, doi:10.1029/2008WR006825, 2009.

Tolstykh, M. A., Diansky, N. A., Gusev, A. V. and Kiktev, D. B.: Simulation of Seasonal Anomalies of Atmospheric Circulation Using Coupled Atmosphere–Ocean Model, Izv. Atmos. Ocean. Phys., 50(2), 131–142, doi:10.7868/S0002351514020126, 2014.

Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E. and Uhlenbrook, S.: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa, Hydrol. Earth Syst. Sci., 19(4), 1695–1711, doi:10.5194/hess-19-1695-2015, 2015.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11, 20131162, 2014.

Weisheimer, A., Doblas-Reyes, F. J., Jung, T. and Palmer, T. N.: On the predictability of the extreme summer 2003 over Europe, Geophys. Res. Lett., 38(5), 1-5, doi:10.1029/2010GL046455, 2011.

Wetterhall, F., Winsemius, H. C., Dutra, E., Werner, M. and Pappenberger, F.: Seasonal predictions of agro-meteorological drought indicators for the Limpopo basin, Hydrol. Earth Syst. Sci., 19, 2577-2586, doi:10.5194/hess-19-2577-2015, 2015.

Wilks, D.S.: Statistical methods in the atmospheric sciences, 3rd edn., Elsevier, 2011.

Wood, A. W., E. P. Maurer, A. Kumar, and D. Lettenmaier, Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res., 107(D20), 4429, doi:10.1029/2001JD000659, 2002.

Zhao, T., Bennett, J., Wang, Q. J., Schepen, A., Wood, A., Robertson, D. and Ramos, M.-H.: How suitable is quantile mapping for post-processing GCM precipitation forecasts?, J. Clim., JCLI-D-16-0652.1, doi:10.1175/JCLI-D-16-0652.1, 2017.

~~8 Figure Captions~~

~~**Figure 1:** Spatio-temporal variability of precipitation, temperature and reference evapotranspiration of monthly aggregated values (P, E_{T0}) and monthly averages (T) of the observation period (1990-2013).~~

~~**Figure 2:** Example of a monthly forecast, valid for August and lead time 1 month for a grid point located in west-central Denmark. Blue shaded box-plots are ensemble climatological forecasts. Black box-plots are ECMWF System 4 raw or postprocessed forecasts. Red dots represent observed values. Bias as in Eq. (4) and Eq. (5) and skill scores for accuracy and sharpness (CRPSS and SS) as in Eq. (8). First column corresponds to the raw forecast, second and third columns correspond to the corrected forecasts with the Linear Scaling/Delta Change (LS) and Quantile Mapping (QM) methods, respectively.~~

Formatted: Line spacing: Multiple 1,15 li, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Check spelling and grammar

Formatted: Line spacing: single, Widow/Orphan control

Formatted: Font: Font color: Black, (none), Check spelling and grammar

Formatted: Space Before: 0 pt, After: 10 pt

~~**Figure 3:** Percentage bias and absolute of monthly values of raw forecasts. Y-axis represents the target month and the X-axis represents the different lead times at which target months are forecasted. Values in blue range represent a positive bias and values in red represent a negative bias.~~

~~**Figure 4:** Percentage bias and absolute bias of monthly values of raw forecasts for the summer. Forecast lead time of 1 month.~~

~~**Figure 5:** Skill in terms of sharpness and accuracy of monthly values of raw forecasts. Y-axis represents the target month and X-axis represents the different lead times at which target month is forecasted. Changes in color represent different skill in accuracy levels and different shapes represent difference in sharpness. Grey color means that the skill in terms of accuracy and squared shape means that the skill in terms of sharpness are respectively worse than the ensemble climatological forecasts.~~

~~**Figure 6:** Spatial variability of skill in accuracy for summer raw forecasts for lead time 1 month. The grids marked with '*' are points where the distribution of the accuracy for ensemble climatology differs from the accuracy distribution of the ECMWF System 4 forecasts at a 5% significance using the WMW-test.~~

~~**Figure 7:** PIT diagrams of summer P, T and E$_{T0}$ for the raw and postprocessed forecasts for lead time 1 month. The interpretation of the PIT diagram is explained in Sect. 5 and Fig. A12.~~

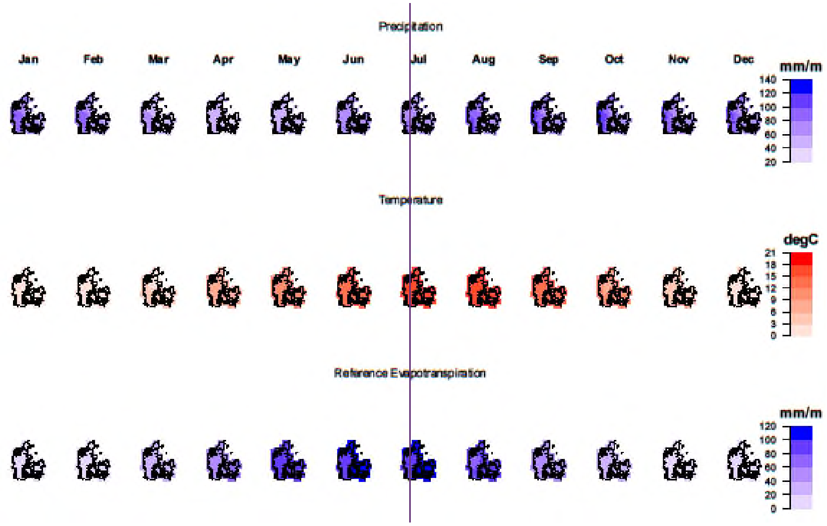
~~**Figure 8:** Biases of raw and postprocessed P, T and E$_{T0}$ at four locations in Denmark. Biases are for the different target months and for lead time 1 month.~~

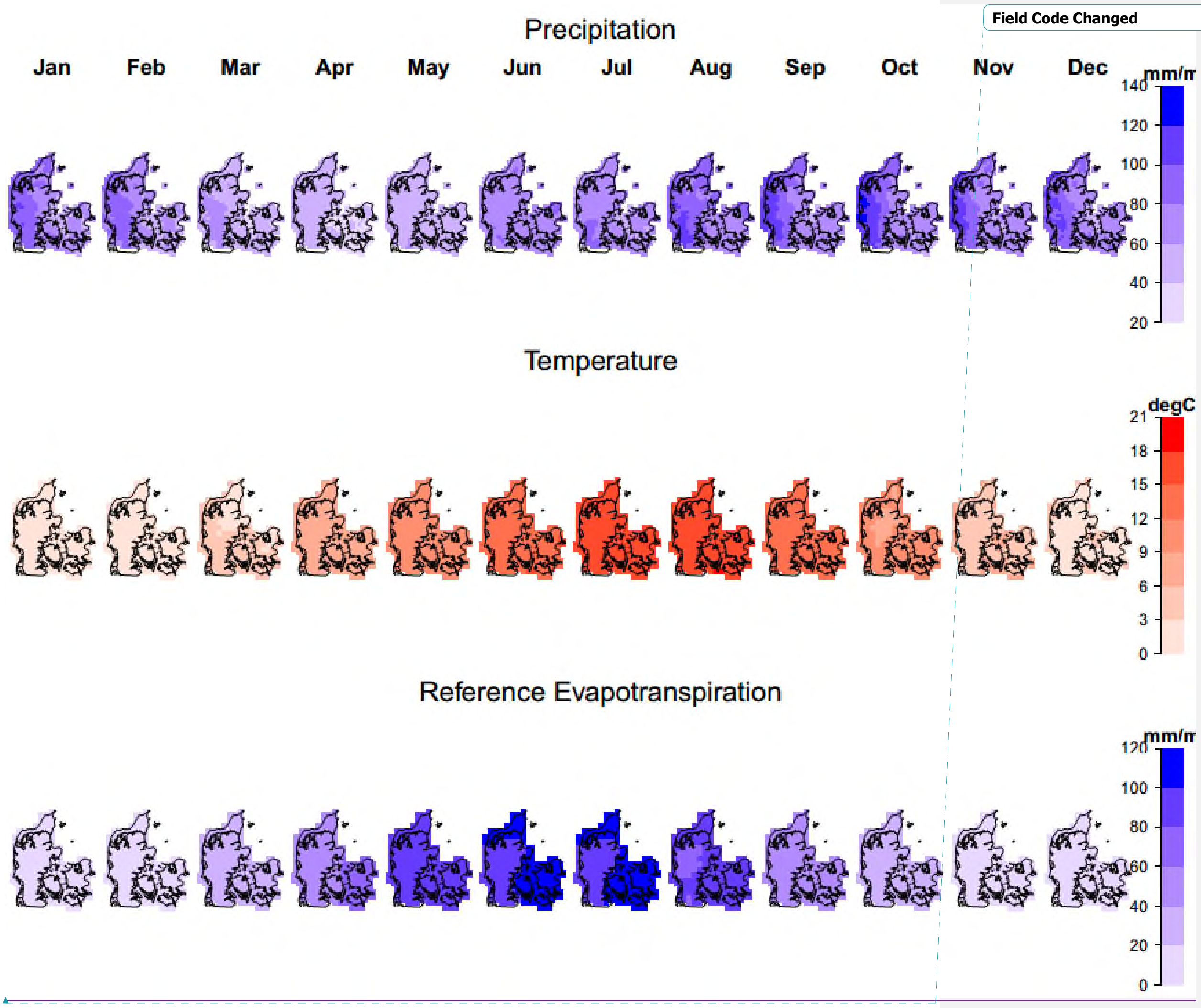
~~**Figure 9:** Percentage of grid points with statistically significant positive CRPSS cf. Eq. (8).~~

~~**Figure 10:** Spatial variability of skill in terms of accuracy for P, T and E$_{T0}$ for the raw and post-processed forecasts of February as a target month at different lead times. Box-plots represent the values of CRPSS cf. Eq. (8) with climatology as reference, of all the 662/724 grid points covering Denmark.~~

~~**Figure 11:** Skill of daily monthly maximum P and number of dry days for target month January for all 7-month lead times for the raw and post-processed forecasts. Box-plots represent the values of CRPSS, of Eq. (8) with climatology as reference, of all the 662/724 grid points covering Denmark.~~

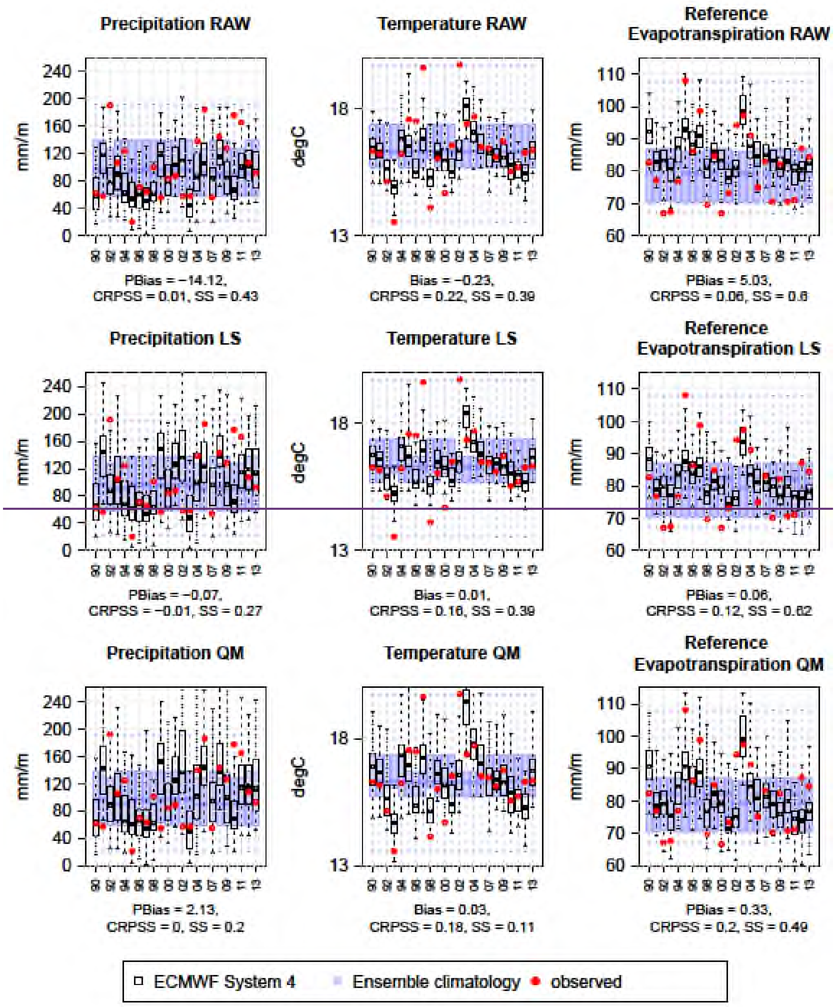
~~**Figure A12:** Probability Integral Transform (PIT) Diagram. Modified from (Laio and Tamea, 2007). Observed values are generated from a standard normal distribution *N*(0,1). Forecasts for these observations are then generated as follows: *N*(1.5,1) for overestimation, *N*(-1.5,1) for underestimation, *N*(0,3) for an overdispersive system and *N*(0.0.3) for a underdispersive system.~~


Precipitation
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
mm/m
140
120
100
80
60
40
20
Temperature
degC
21
18
15
12
9
6
3
0
Reference Evapotranspiration
mm/m
120
100
80
60
40
20
0



Field Code Changed

Figure 1: Spatio-temporal variability of precipitation, temperature and reference evapotranspiration of monthly aggregated values (P, E_{T0}) and monthly averages (T) of the observation period (1990-2013).


Precipitation RAW
Temperature RAW
Reference Evapotranspiration RAW
mm/m
degC
240
200
160
120
80
40
0
18
13
110
100
90
80
70
60
PBias = −14.12, CRPSS = 0.01, SS = 0.43
Bias = −0.23, CRPSS = 0.22, SS = 0.39
PBias = 5.03, CRPSS = 0.06, SS = 0.6
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
PBias = −0.07, CRPSS = −0.01, SS = 0.27
Bias = 0.01, CRPSS = 0.16, SS = 0.39
PBias = 0.06, CRPSS = 0.12, SS = 0.62
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
PBias = 2.13, CRPSS = 0, SS = 0.2
Bias = 0.03, CRPSS = 0.18, SS = 0.11
PBias = 0.33, CRPSS = 0.2, SS = 0.49
□ ECMWF System 4
Ensemble climatology
● observed

Field Code Changed

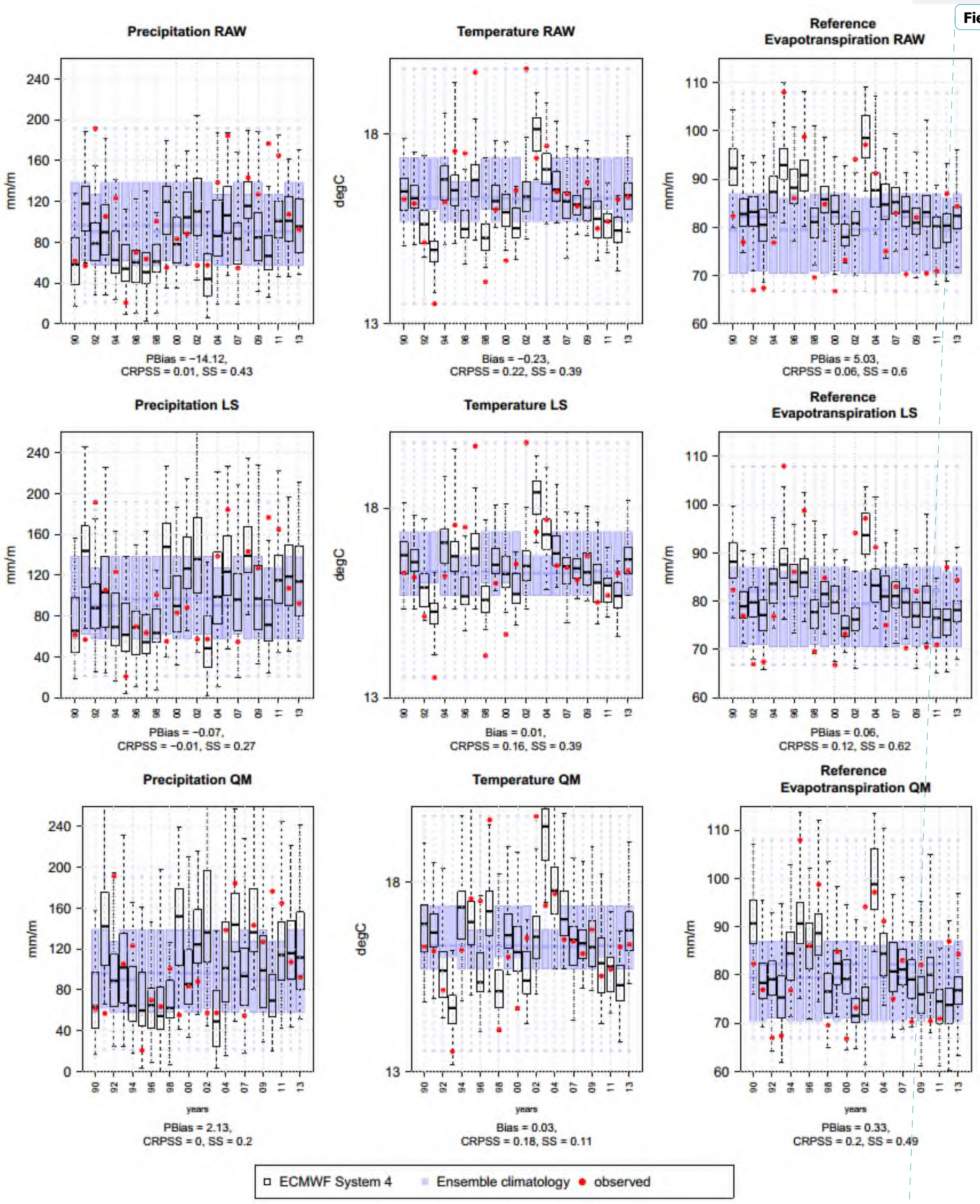
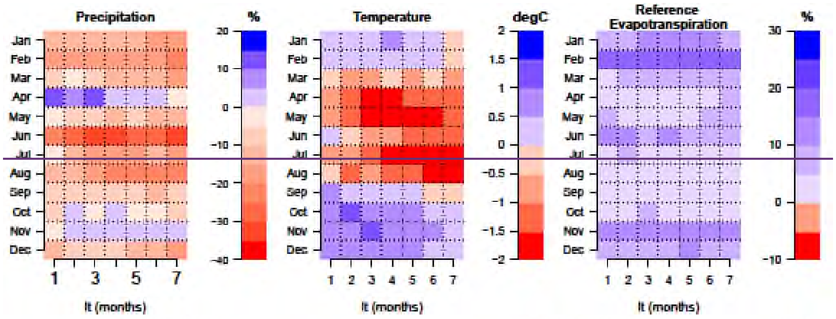
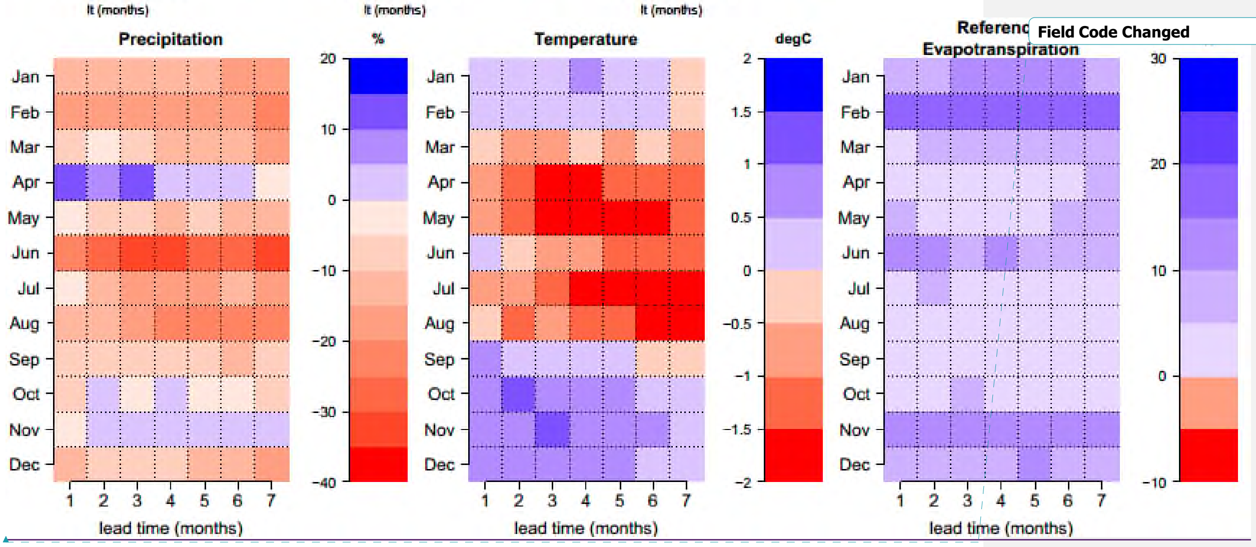


Figure 2: Example of a monthly forecast, valid for August and lead time 1 month for a grid point located in west-central Denmark. Blue ~~shaded~~ box-plots are the ensemble climatological forecasts. Black box-plots are ECMWF System 4 raw or

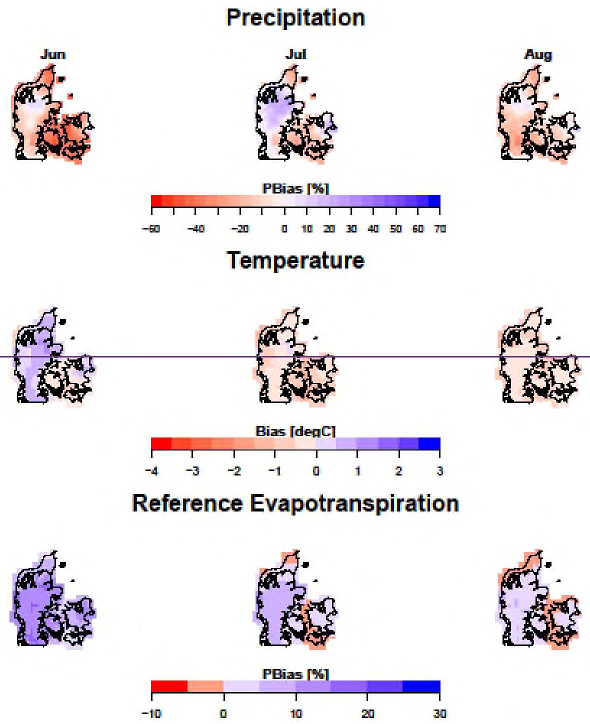
postprocessed forecasts. Red dots represent observed values. Bias as in Eq. (4) and Eq. (5) and skill scores for accuracy and sharpness (CRPSS and SS) as in Eq. (8). First column corresponds to the raw forecast, second and third columns correspond to the corrected forecasts with the Linear Scaling/Delta Change (LS) and Quantile Mapping (QM) methods, respectively.

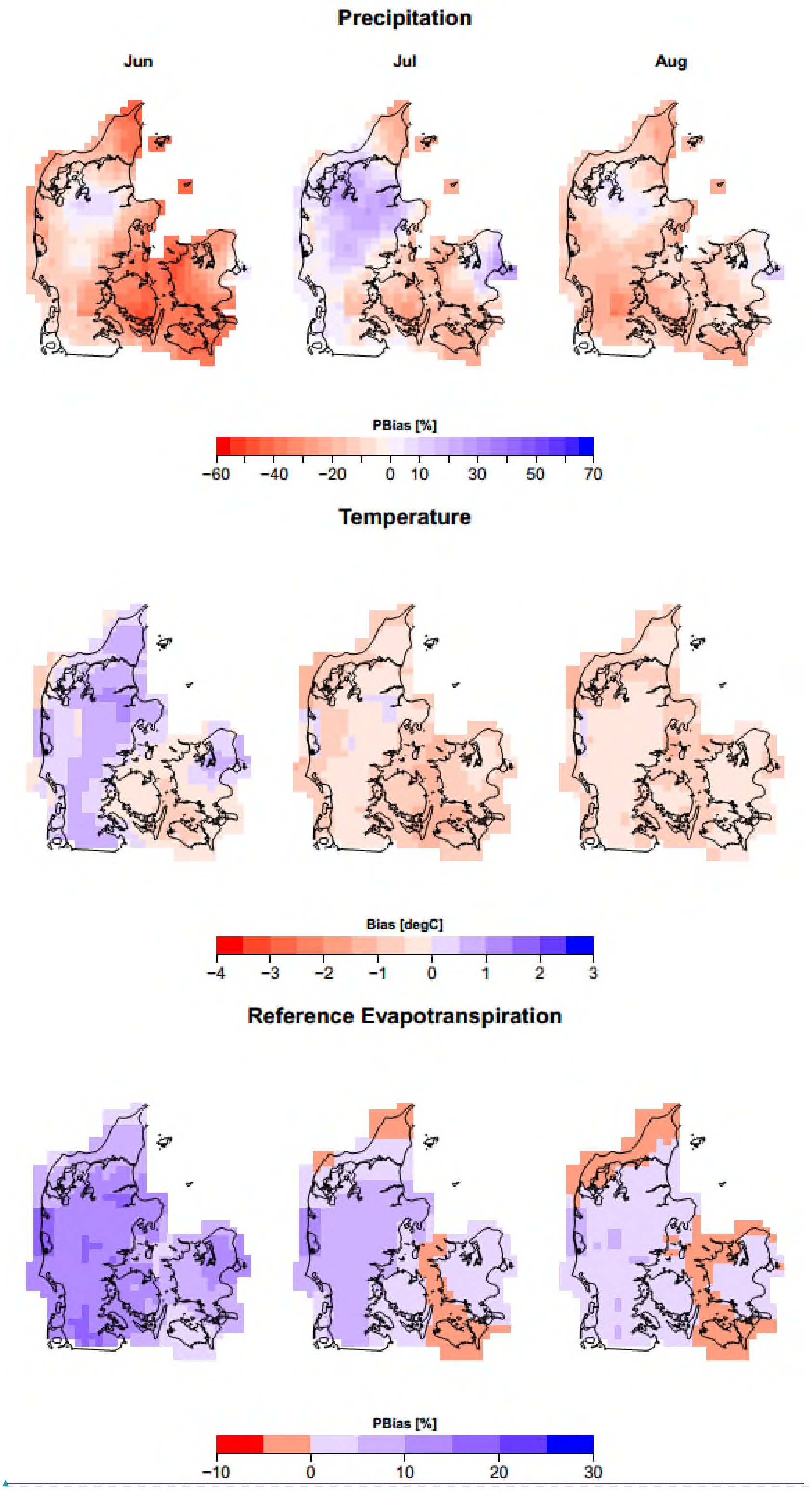




Field Code Changed

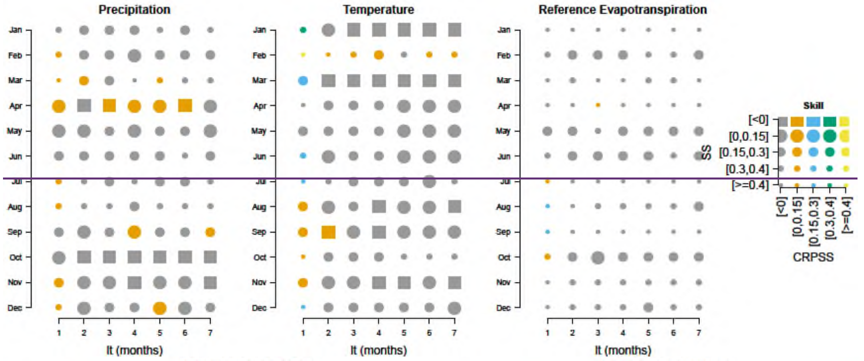
Figure 3: Percentage bias and absolute of monthly values of raw forecasts. Y-axis represents the target month and the X-axis represents the different lead times at which target months are forecasted. Values in blue range represent a positive bias and values in red represent a negative bias.


Precipitation
Jun
Jul
Aug
PBias [%]
−60
−40
−20
0
10
20
30
40
50
60
70
Temperature
Bias [degC]
−4
−3
−2
−1
0
1
2
3
Reference Evapotranspiration
PBias [%]
−10
0
10
20
30



Field Code Changed

Figure 4: Percentage bias and absolute bias of monthly values of raw forecasts for the summer. Forecast lead time of 1 month.



Field Code Changed

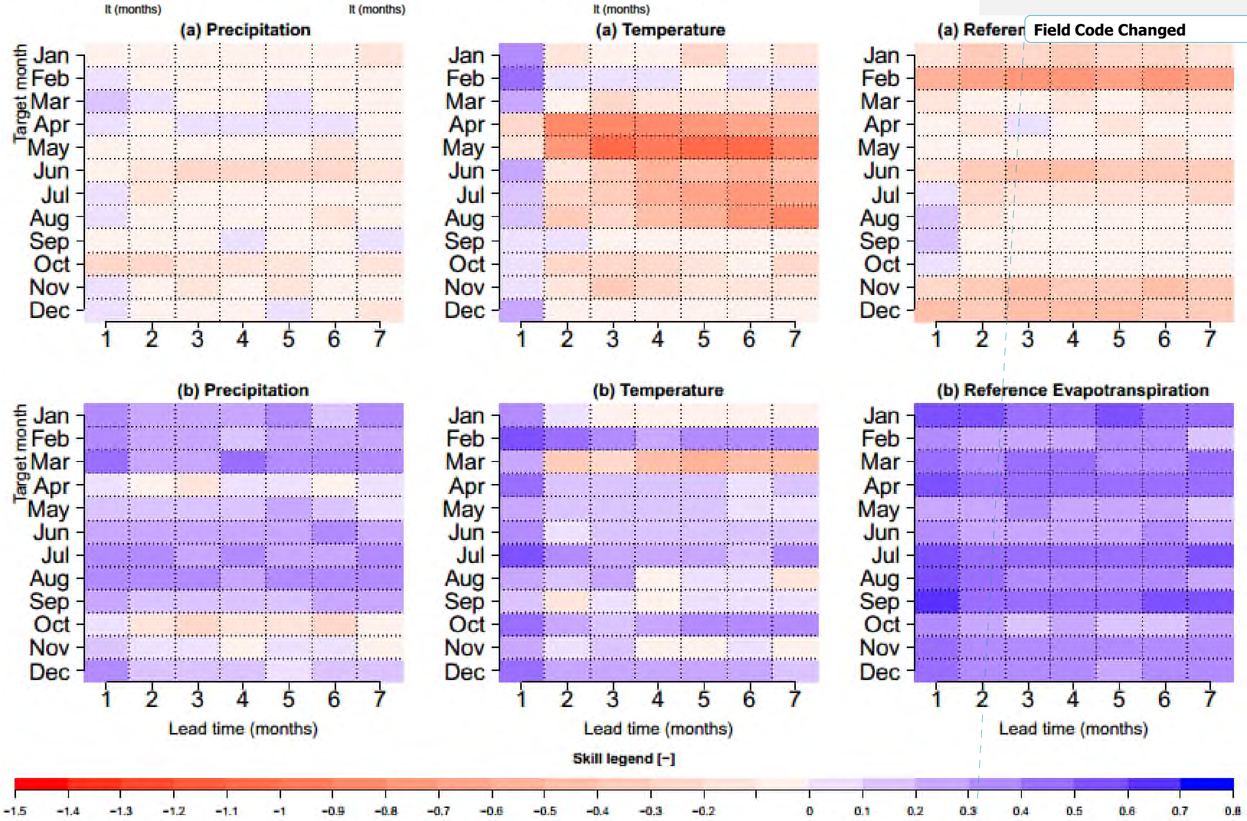
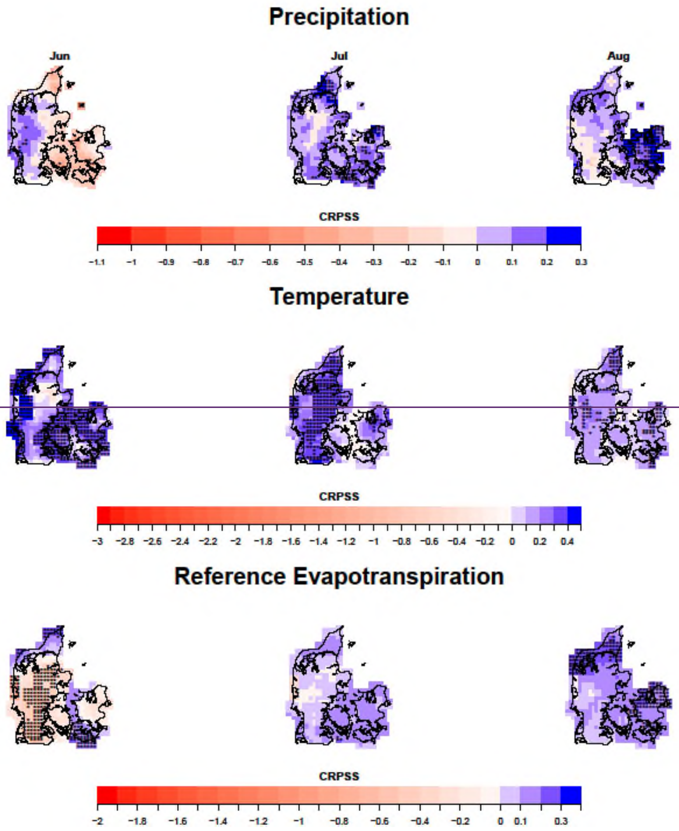
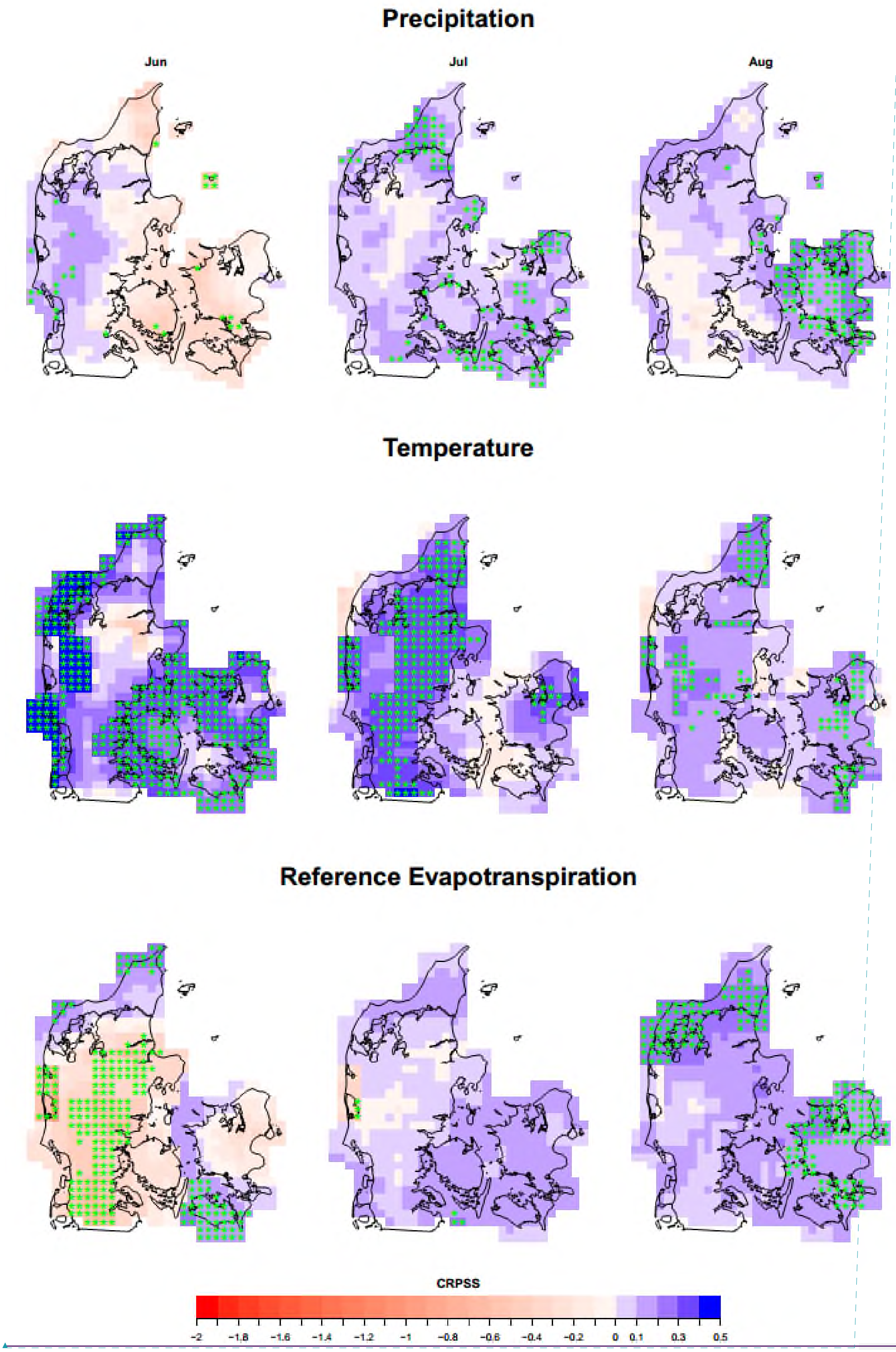


Figure 5: Skill in terms of ~~sharpness~~ accuracy (a) and ~~accuracy~~ sharpness (b) of monthly values of raw forecasts. Y-axis represents the target month and X-axis represents the different lead times at which target month is forecasted. ~~Changes in color represent different skill in accuracy levels and different shapes represent difference in sharpness. Grey color means that the skill in terms of accuracy and squared shape means that the skill in terms of sharpness are respectively worse than the ensemble climatological forecasts.~~

5


Precipitation
Jun
Jul
Aug
CRPSS
−1.1 −1 −0.9 −0.8 −0.7 −0.6 −0.5 −0.4 −0.3 −0.2 −0.1 0 0.1 0.2 0.3
Temperature
CRPSS
−3 −2.8 −2.6 −2.4 −2.2 −2 −1.8 −1.6 −1.4 −1.2 −1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4
Reference Evapotranspiration
CRPSS
−2 −1.8 −1.6 −1.4 −1.2 −1 −0.8 −0.6 −0.4 −0.2 0 0.1 0.3

Field Code Changed

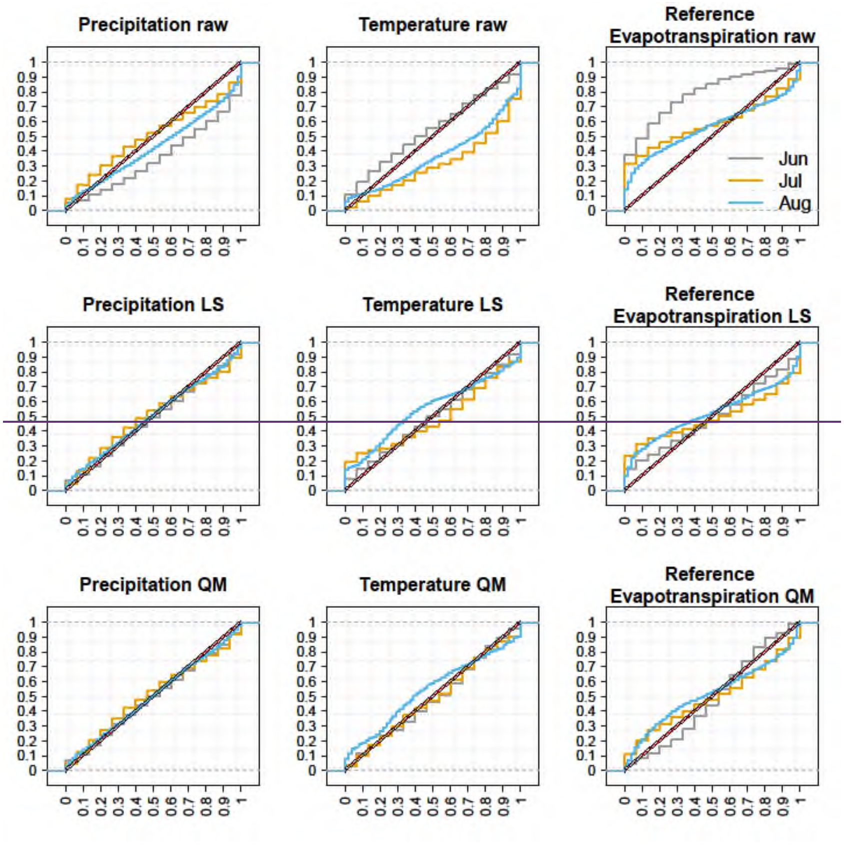


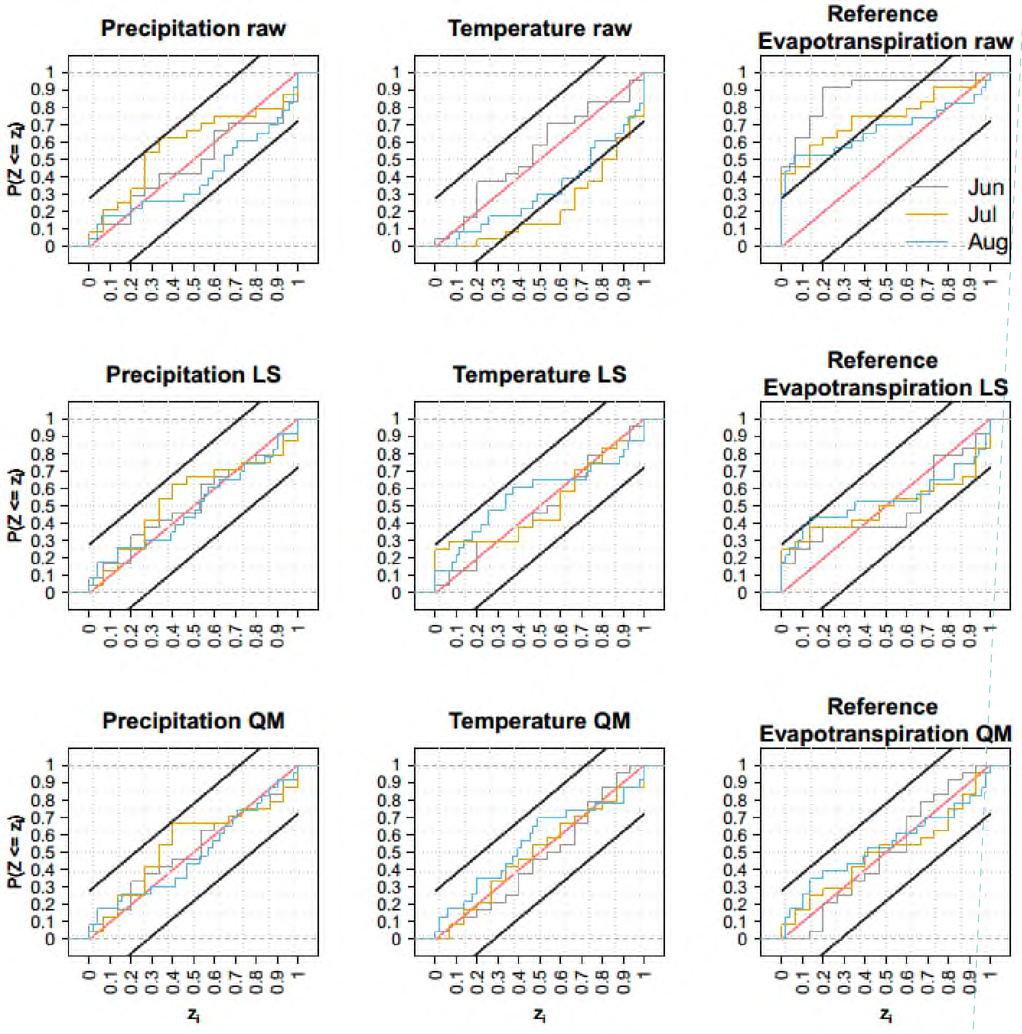
Precipitation

Temperature

Reference Evapotranspiration

Figure 6: Spatial variability of skill in accuracy for summer raw forecasts for lead time 1 month. The grids marked with '*' are points where the distribution of the accuracy for ensemble climatology differs from the accuracy distribution of the ECMWF System 4 forecasts at a 5% significance using the WMW-test.


Precipitation raw
Temperature raw
Reference Evapotranspiration raw
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Jun
Jul
Aug



Field Code Changed

Figure 7: PIT diagrams of summer P, T and E$_{T0}$ for the raw and postprocessed forecasts for lead time 1 month for one grid point located in western Denmark. ~~The interpretation of the PIT diagram is explained in Sect. 5 and Fig. A12.~~

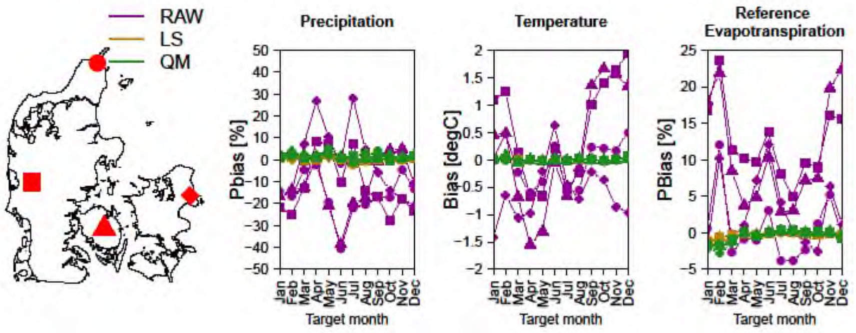
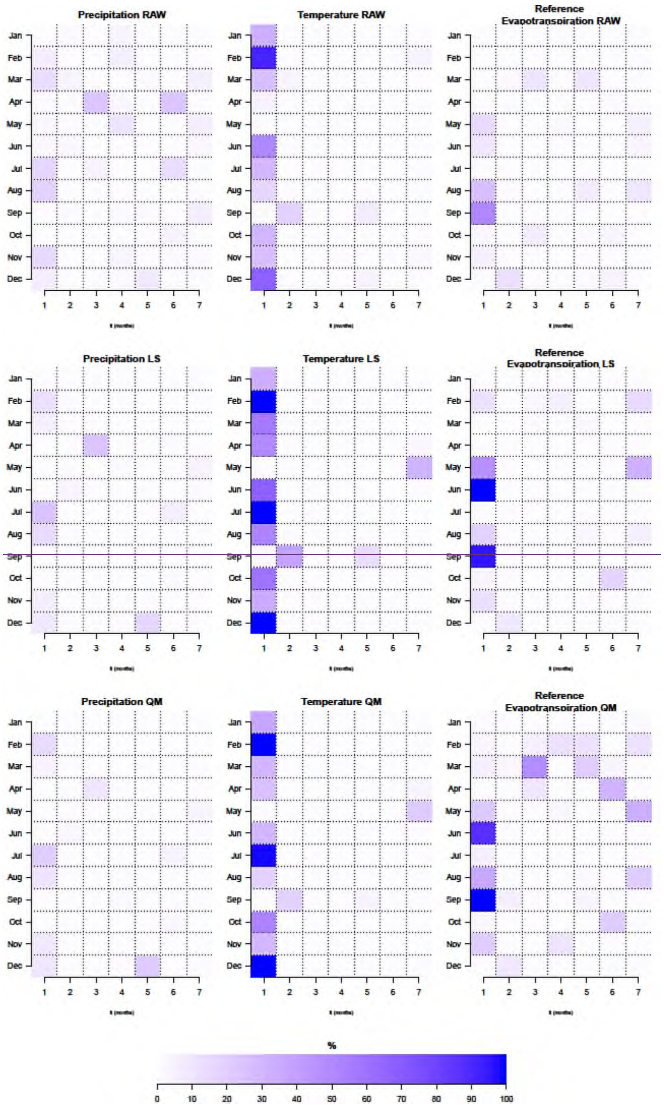


Figure 8: Biases of raw and postprocessed P, T and E_{T0} at four locations in Denmark. Biases are for the different target months and for lead time 1 month. Different locations are represented with different symbol shape according to the map on the left, whereas the raw and the different postprocessing techniques are represented with different colors.


Precipitation RAW
Temperature RAW
Reference Evapotranspiration RAW
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
2
3
4
5
6
7
k (months)
%
0
10
20
30
40
50
60
70
80
90
100

Field Code Changed

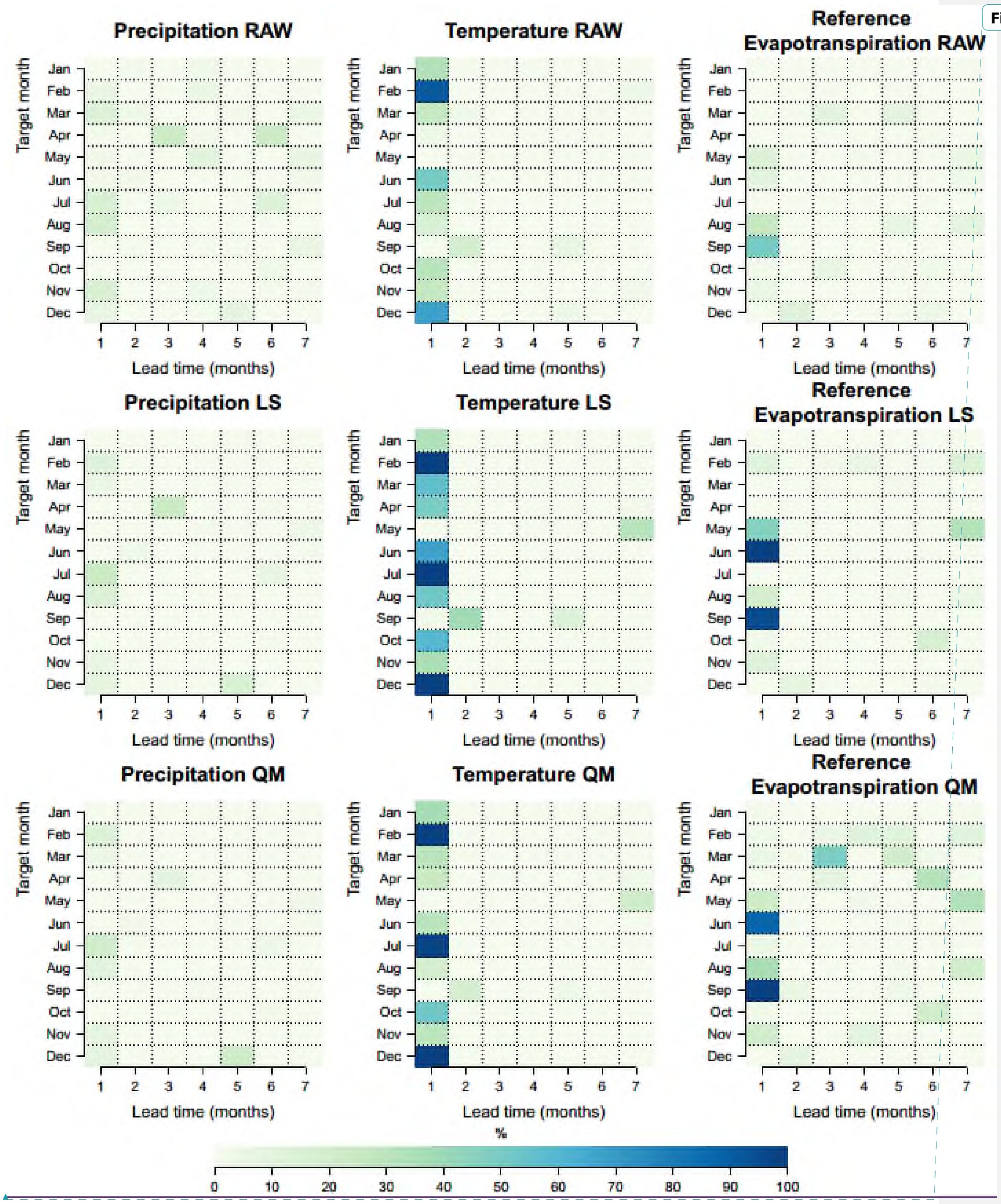
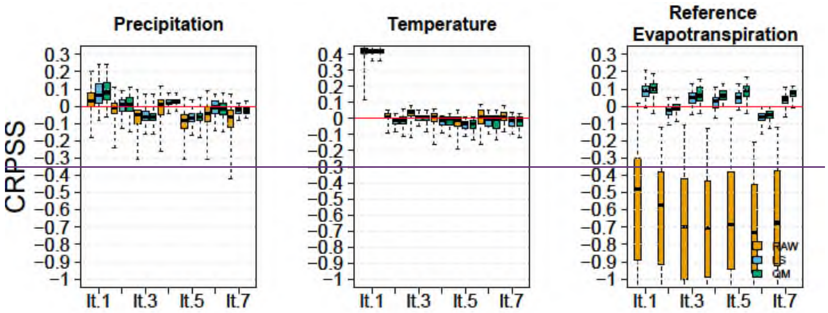
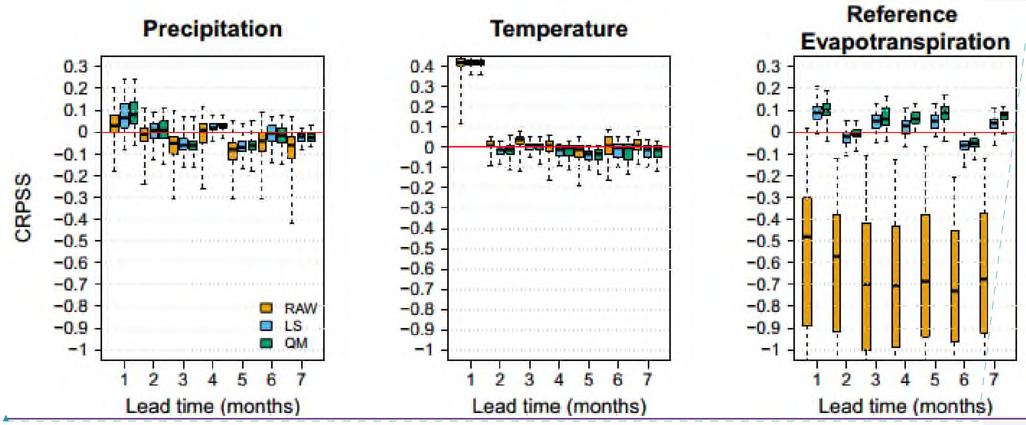


Figure 9: Percentage of grid points with statistically significant positive CRPSS cf. Eq. (8).

5





Field Code Changed

Figure 10: Spatial variability of skill in terms of accuracy for P, T and E_{T0} for the raw and post-processed forecasts of February as a target month at different lead times. Box-plots represent the values of CRPSS cf. Eq. (8) with climatology as reference, of all the 662/724 grid points covering Denmark.

Field Code Changed

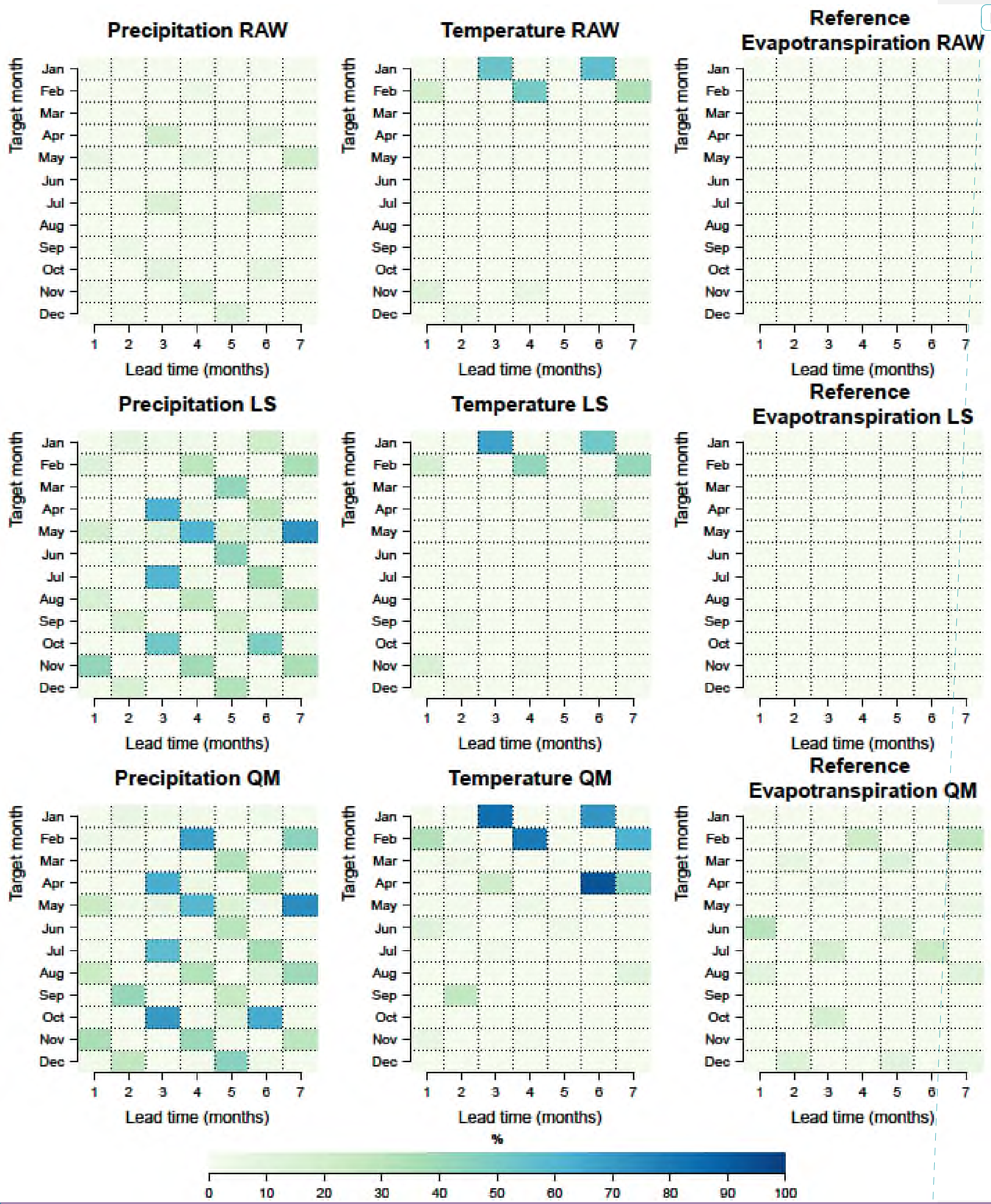
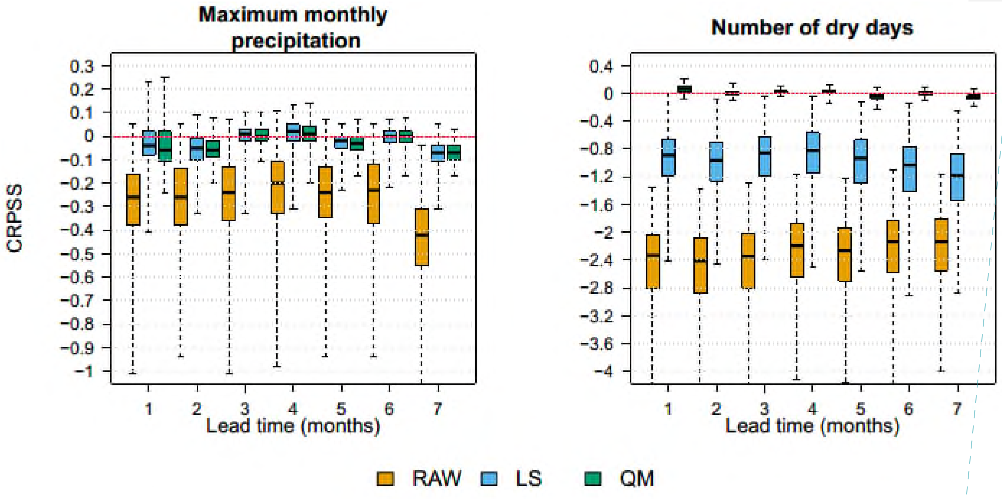
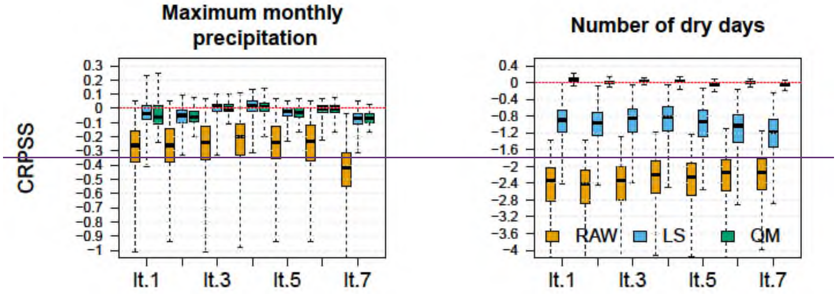


Figure 11: Percentage of grid points for which we fail to reject the null hypothesis of uniformity using the Anderson-Darling test at the 5% significance level.

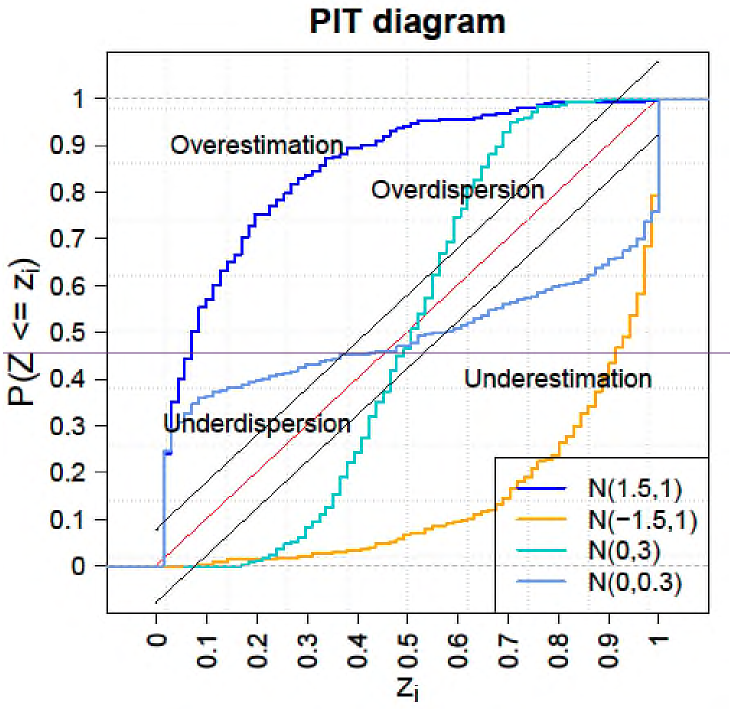
5



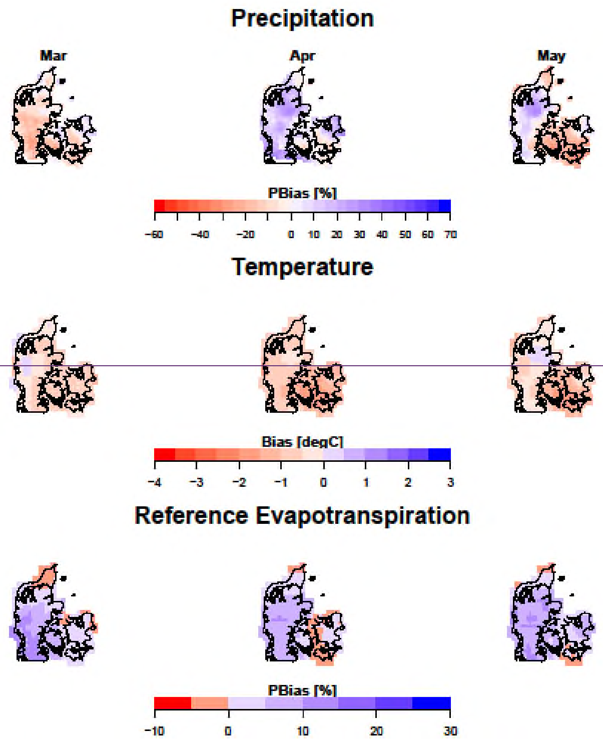
Field Code Changed

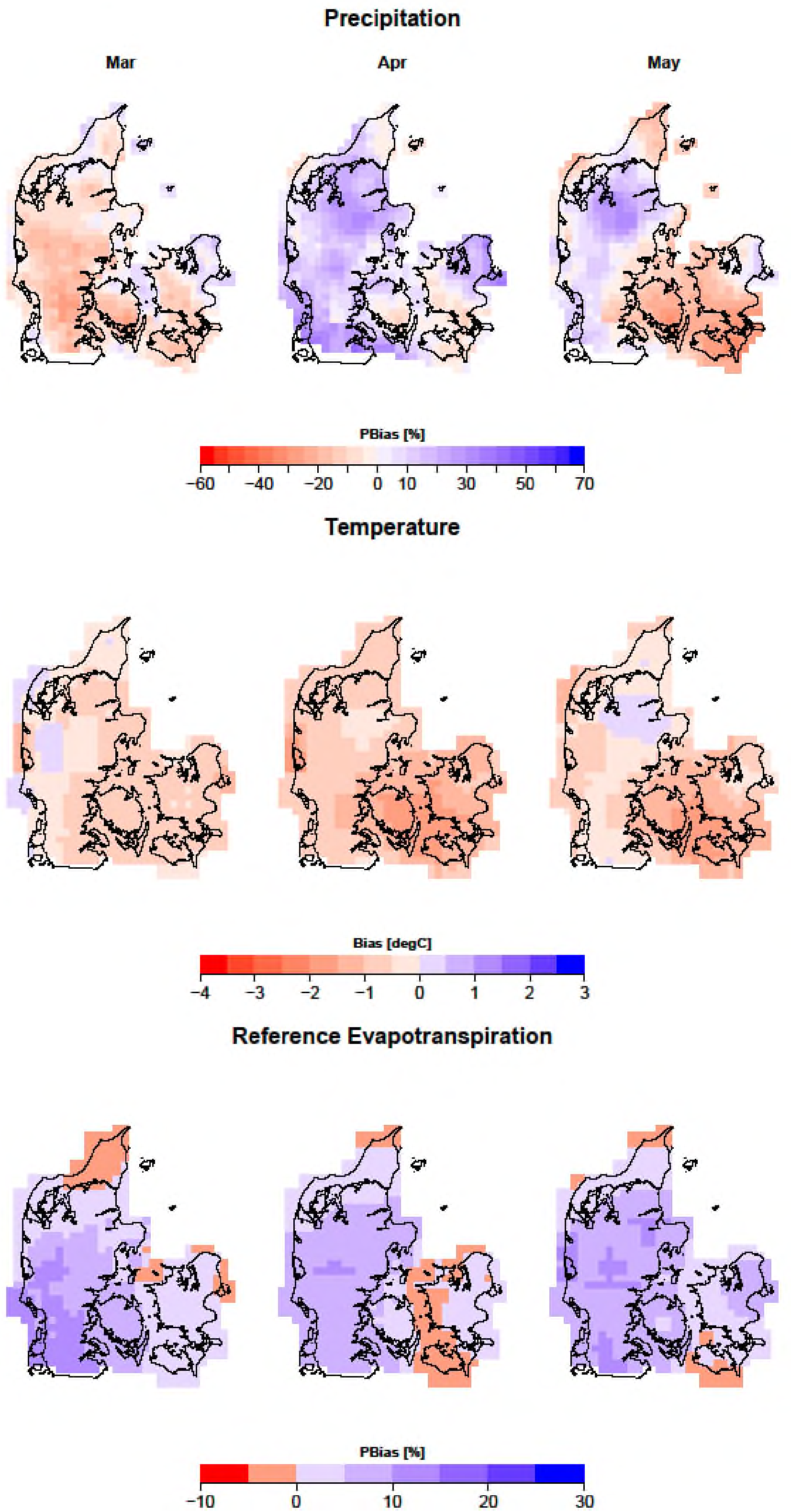


-Figure 12~~1~~: Skill of daily monthly maximum P and number of dry days for target month January for all 7-month lead times for the raw and post-processed forecasts. Box-plots represent the values of CRPSS, of Eq. (8) with climatology as reference, of all the 662/724 grid points covering Denmark.



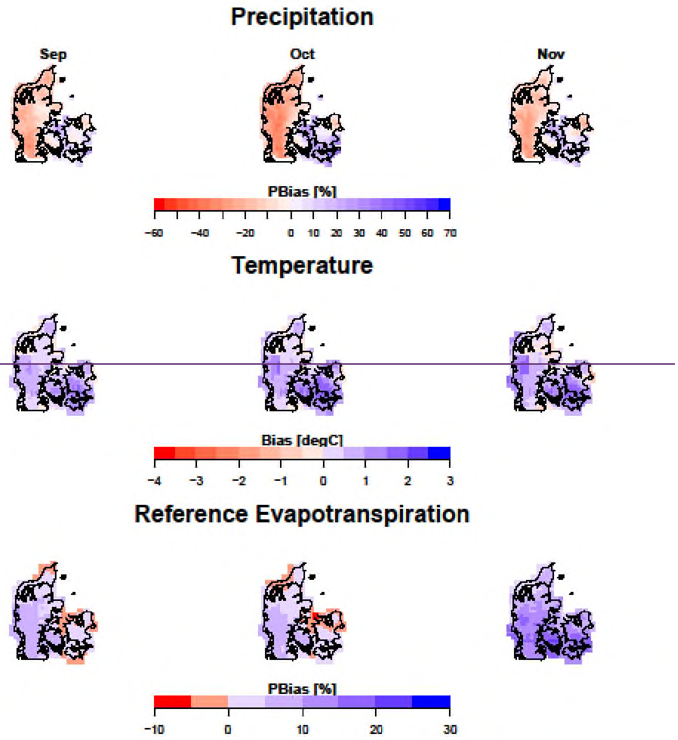
~~**Figure A12:** Probability Integral Transform (PIT) Diagram. Modified from (Laio and Tamea, 2007). Observed values are generated from a standard normal distribution *N*(0,1). Forecasts for these observations are then generated as follows: *N*(1.5,1) for overestimation, *N*(-1.5,1) for underestimation, *N*(0,3) for an overdispersive system and *N*(0.0.3) for a underdispersive system.~~

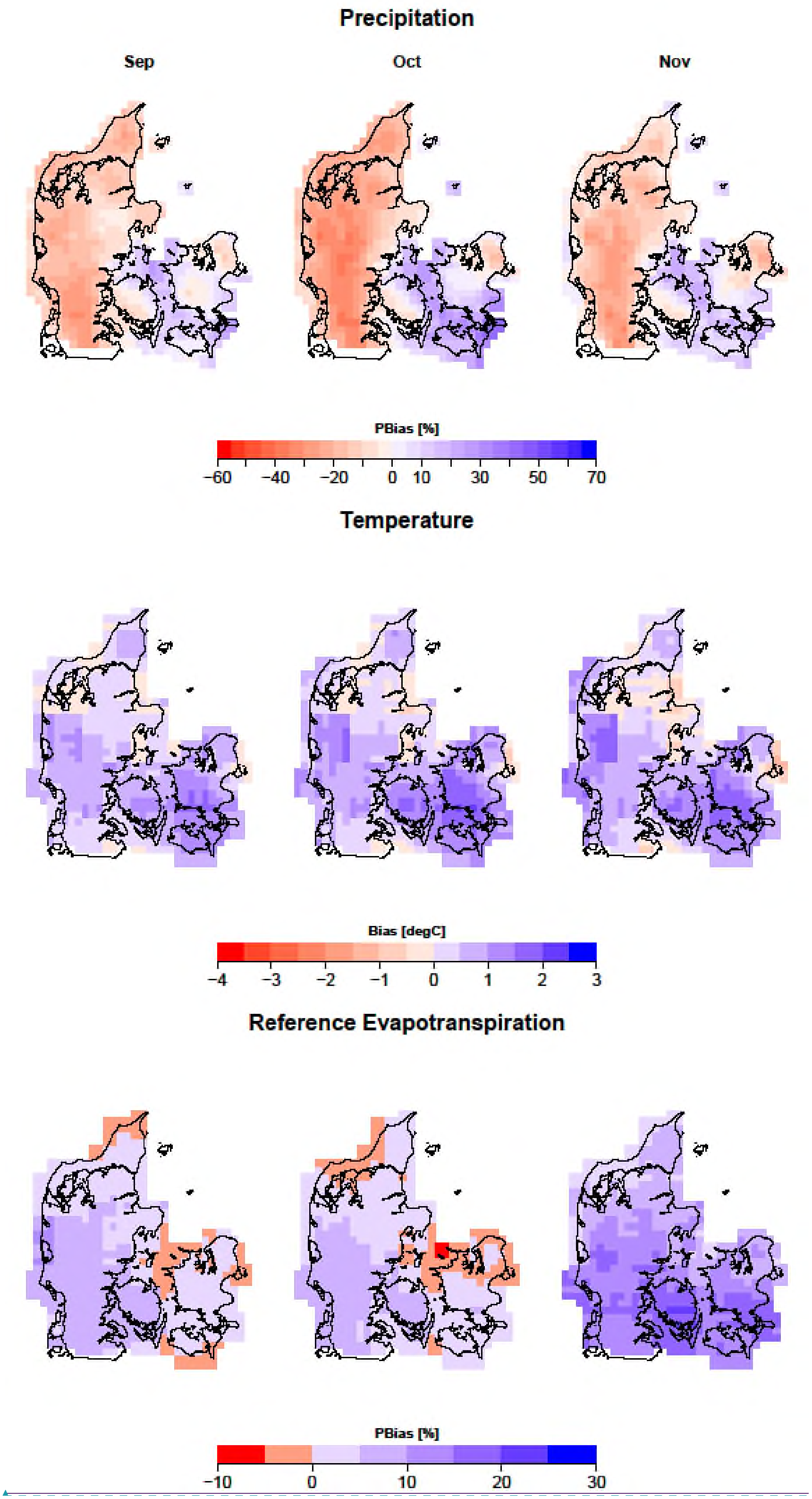

Precipitation
Mar
Apr
May
PBias [%]
−60
−40
−20
0
10
20
30
40
50
60
70
Temperature
Bias [degC]
−4
−3
−2
−1
0
1
2
3
Reference Evapotranspiration
PBias [%]
−10
0
10
20
30



Field Code Changed

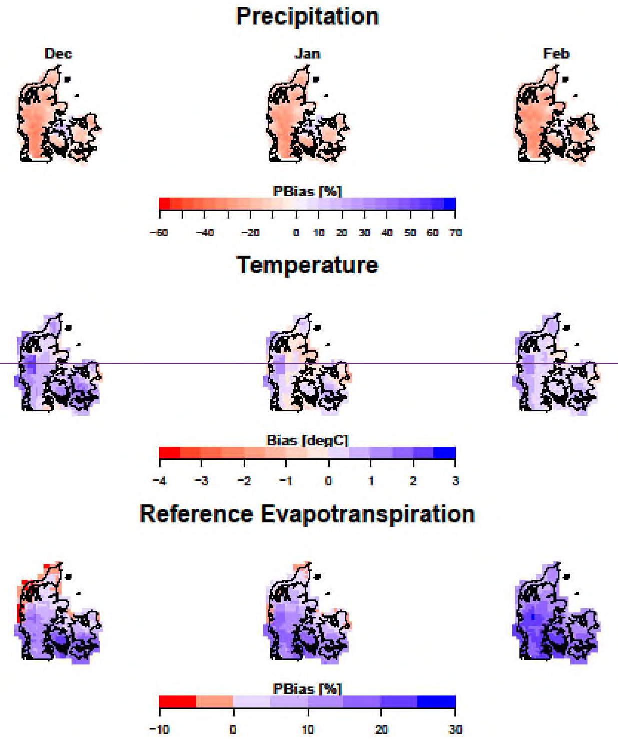
Figure S1: As in Fig. 4, but for spring.

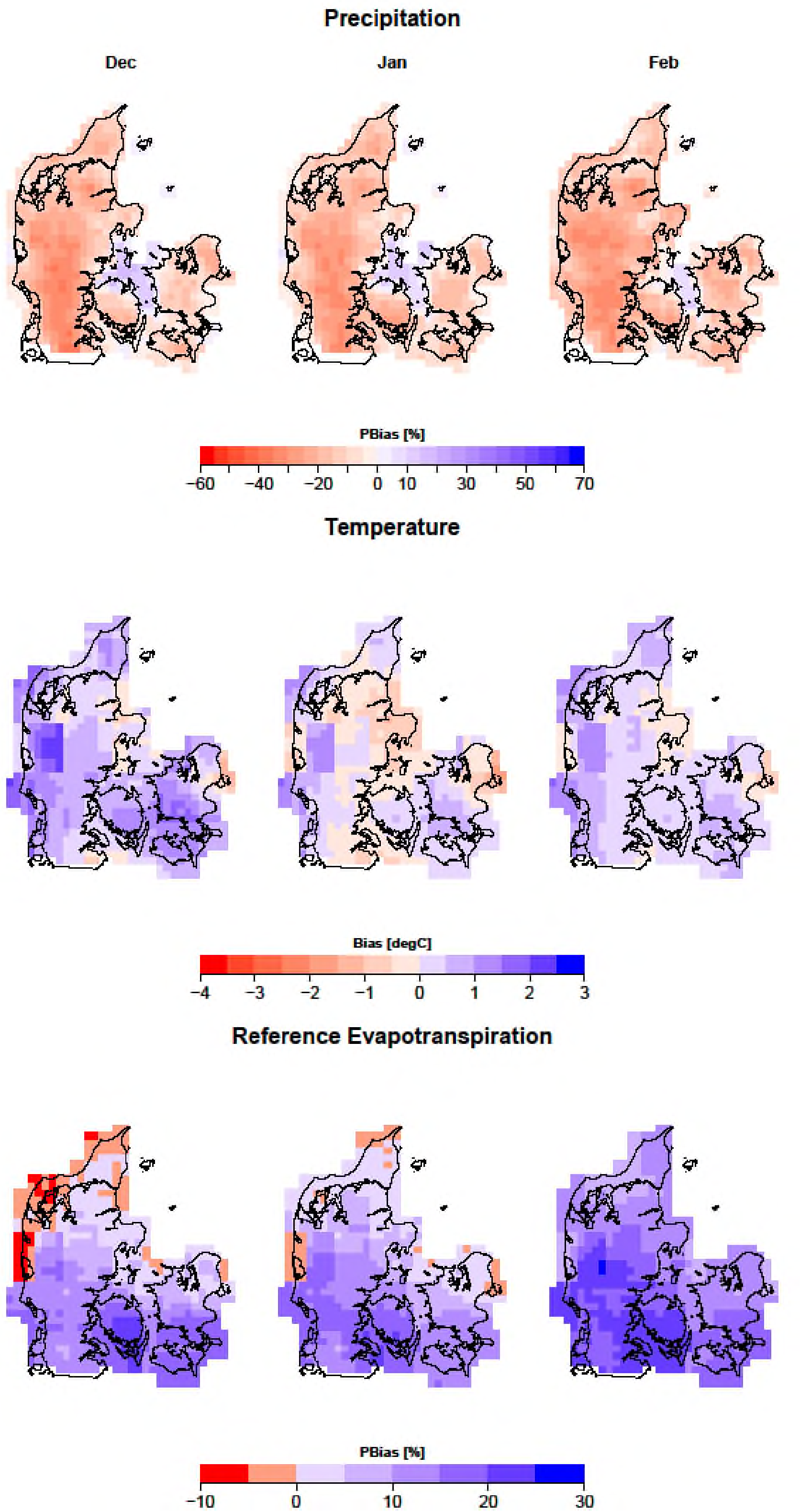




Field Code Changed

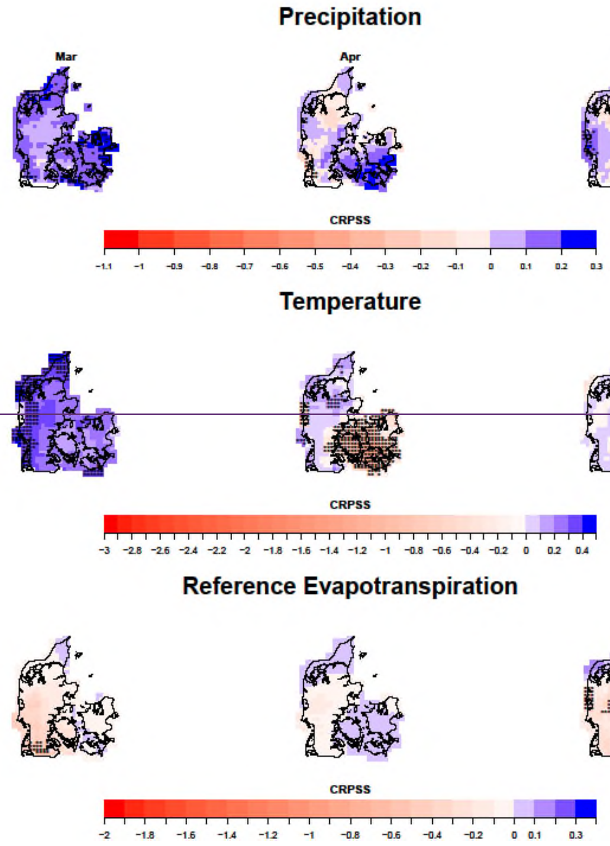
Figure S2: As in Fig. 4, but for autumn.





Field Code Changed

Figure S3: As in Fig. 4, but for winter.



Precipitation

Temperature

Reference Evapotranspiration

Field Code Changed

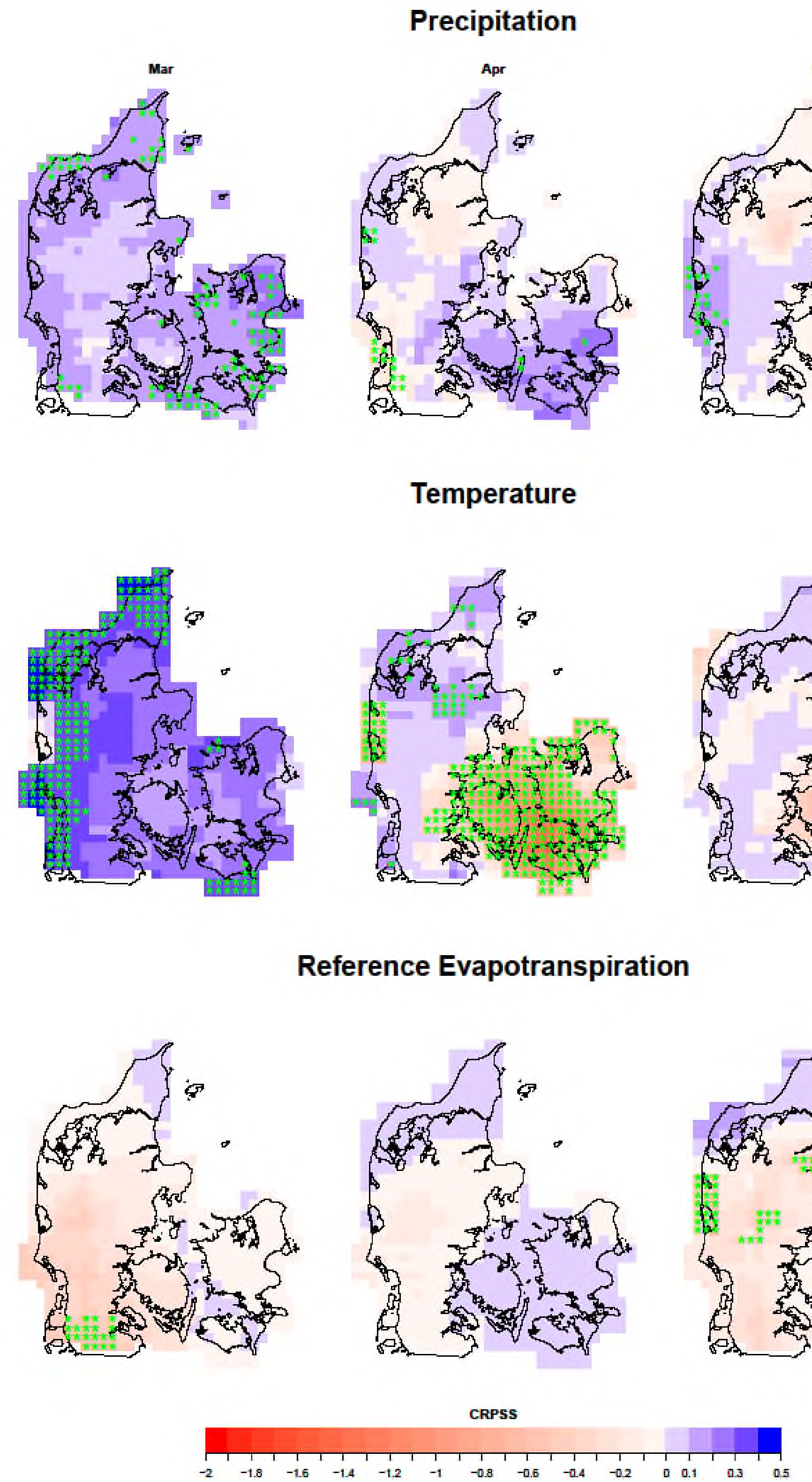
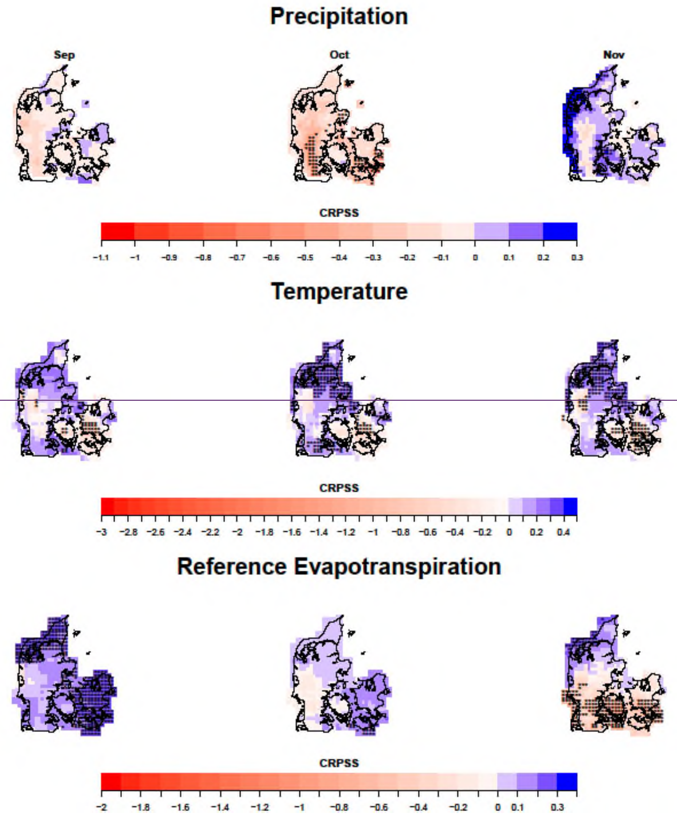


Figure S4: As in Fig. 6, but for spring.



Precipitation
Sep
Oct
Nov
CRPSS
−1.1
−1
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
Temperature
CRPSS
−3
−2.8
−2.6
−2.4
−2.2
−2
−1.8
−1.6
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
Reference Evapotranspiration
CRPSS
−2
−1.8
−1.6
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.1
0.3

Field Code Changed

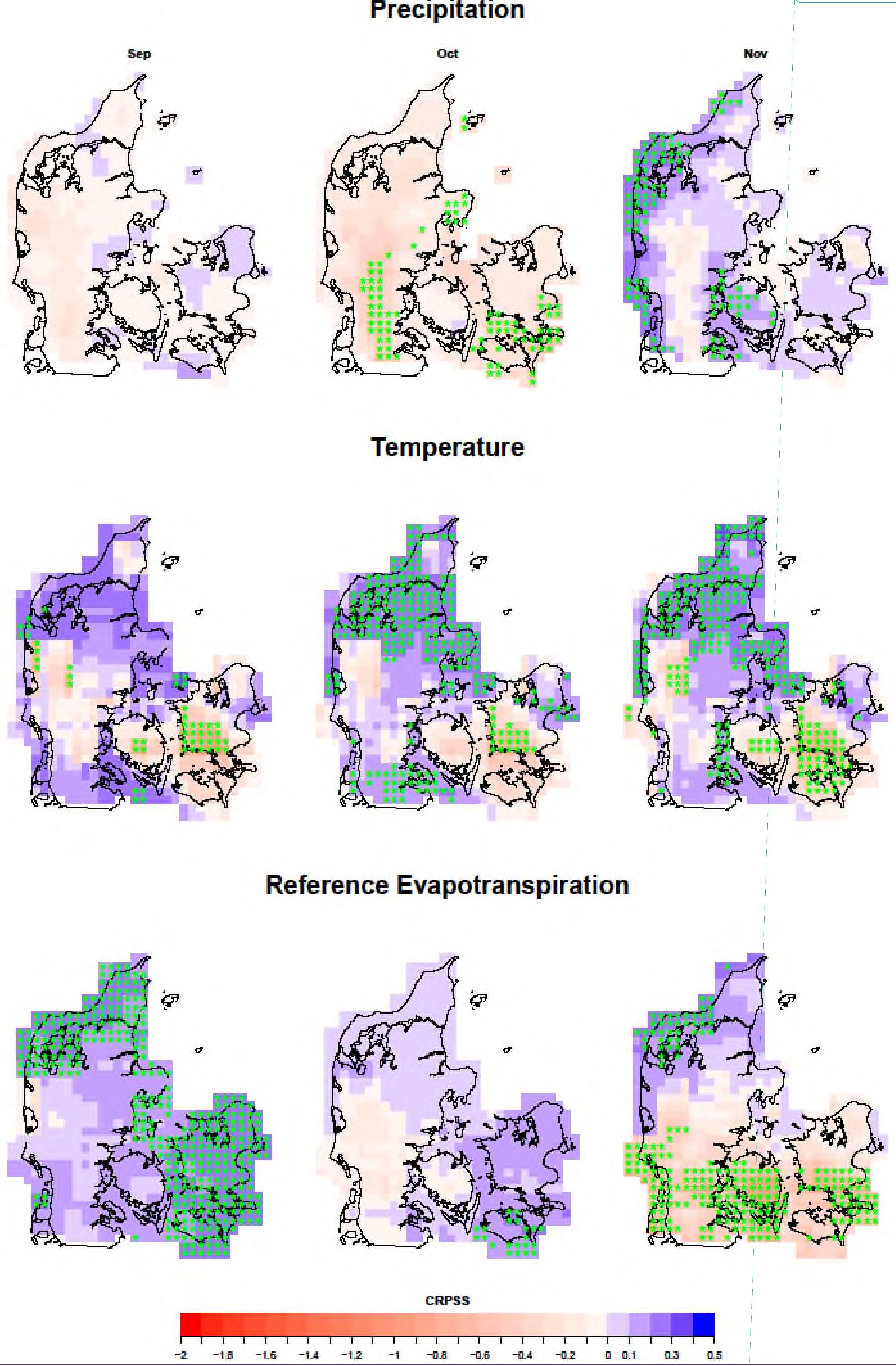
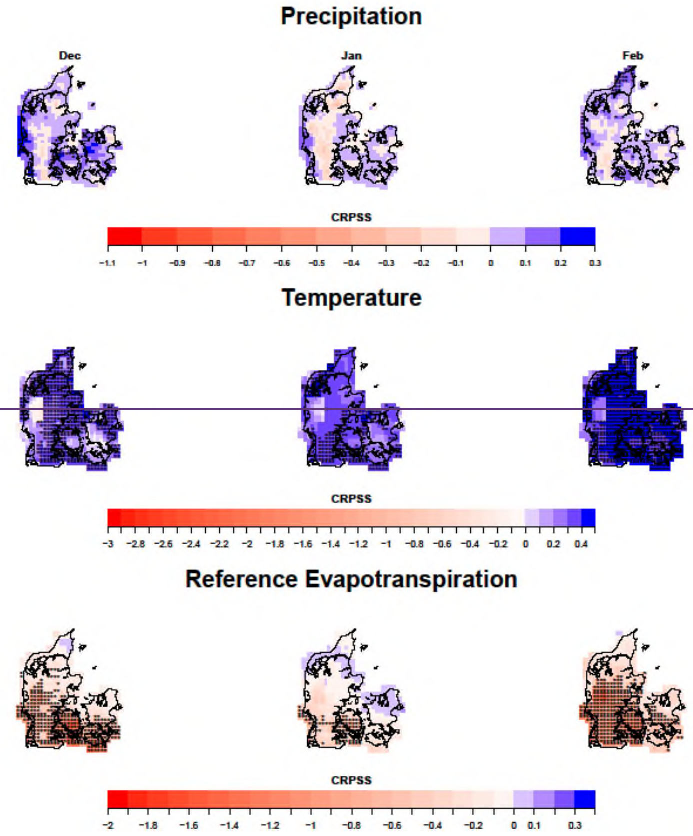


Figure S5: As in Fig. 6, but for autumn.



Precipitation

Temperature

Reference Evapotranspiration

Field Code Changed

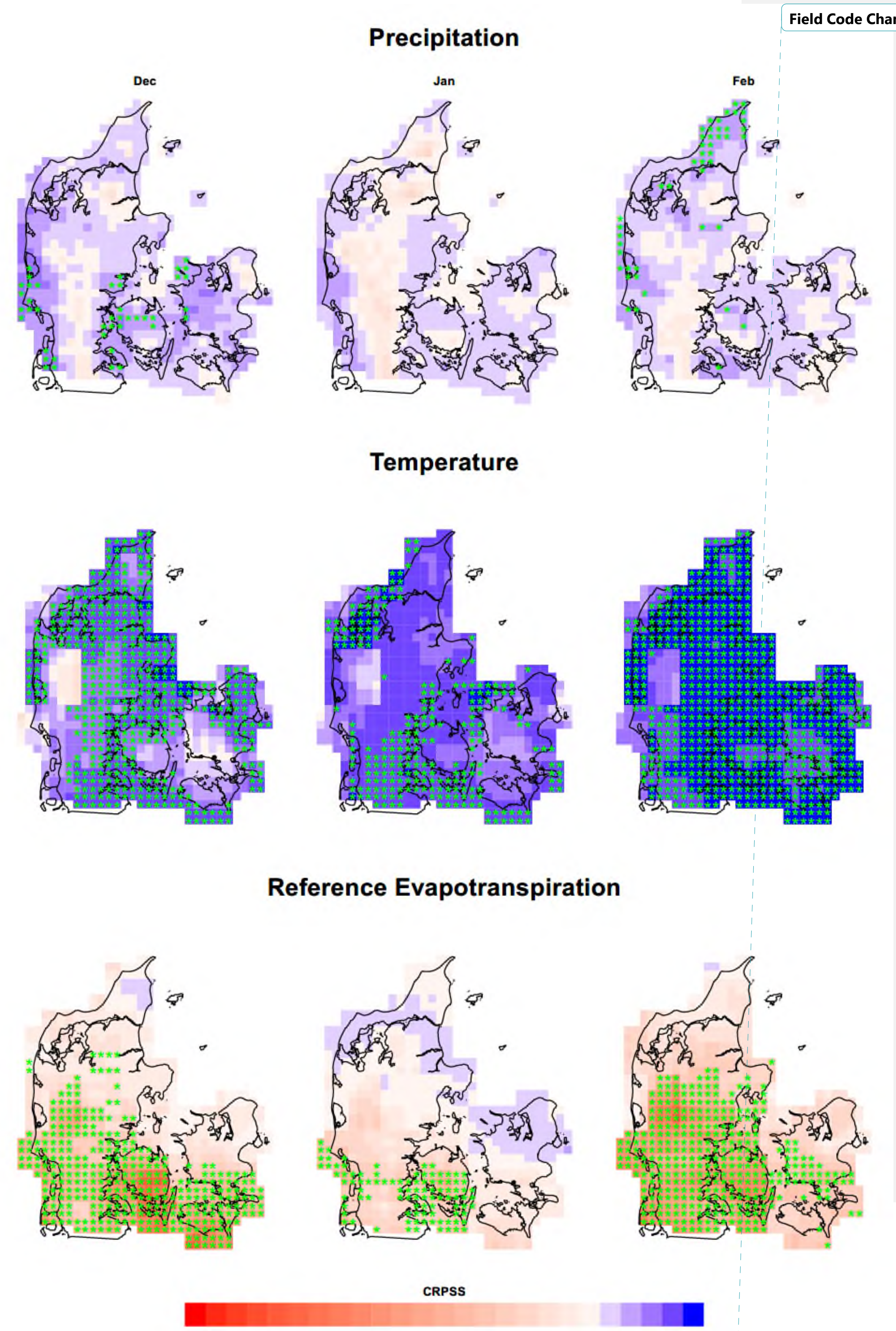
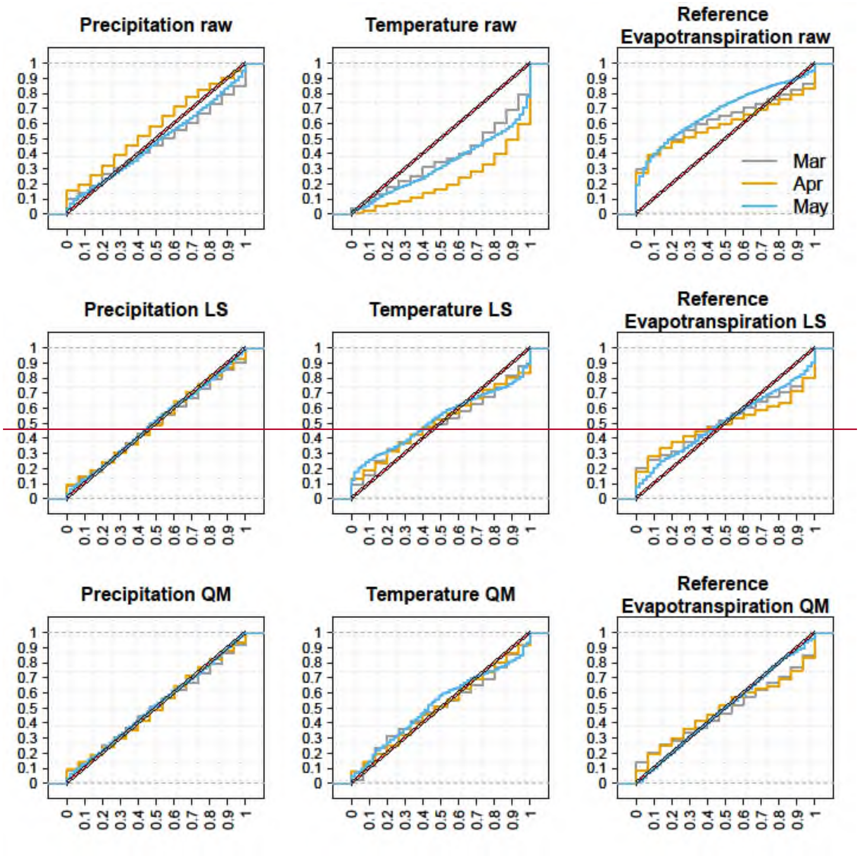


Figure S6: As in Fig. 6, but for winter.

Precipitation raw
Temperature raw
Reference Evapotranspiration raw
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Mar
Apr
May
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Field Code Changed

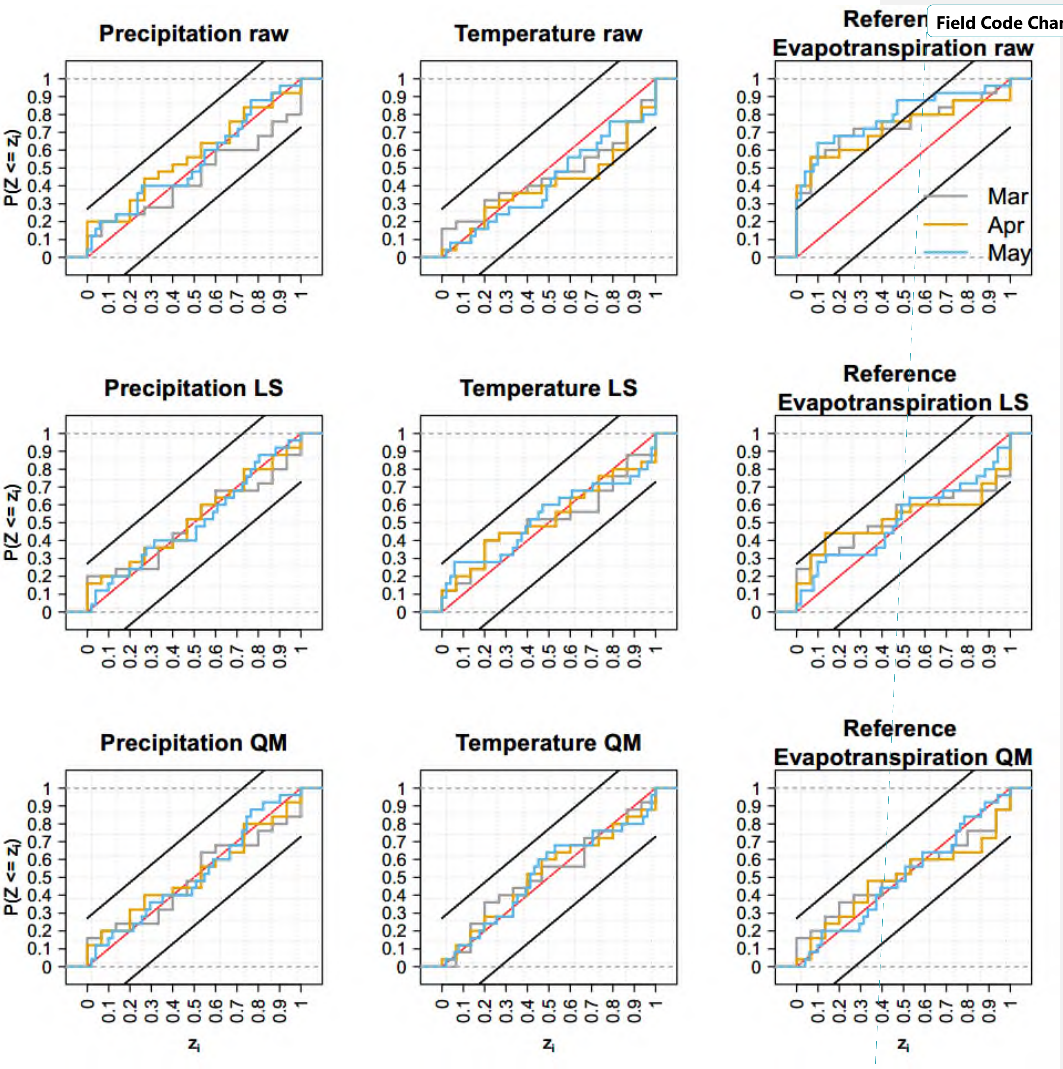
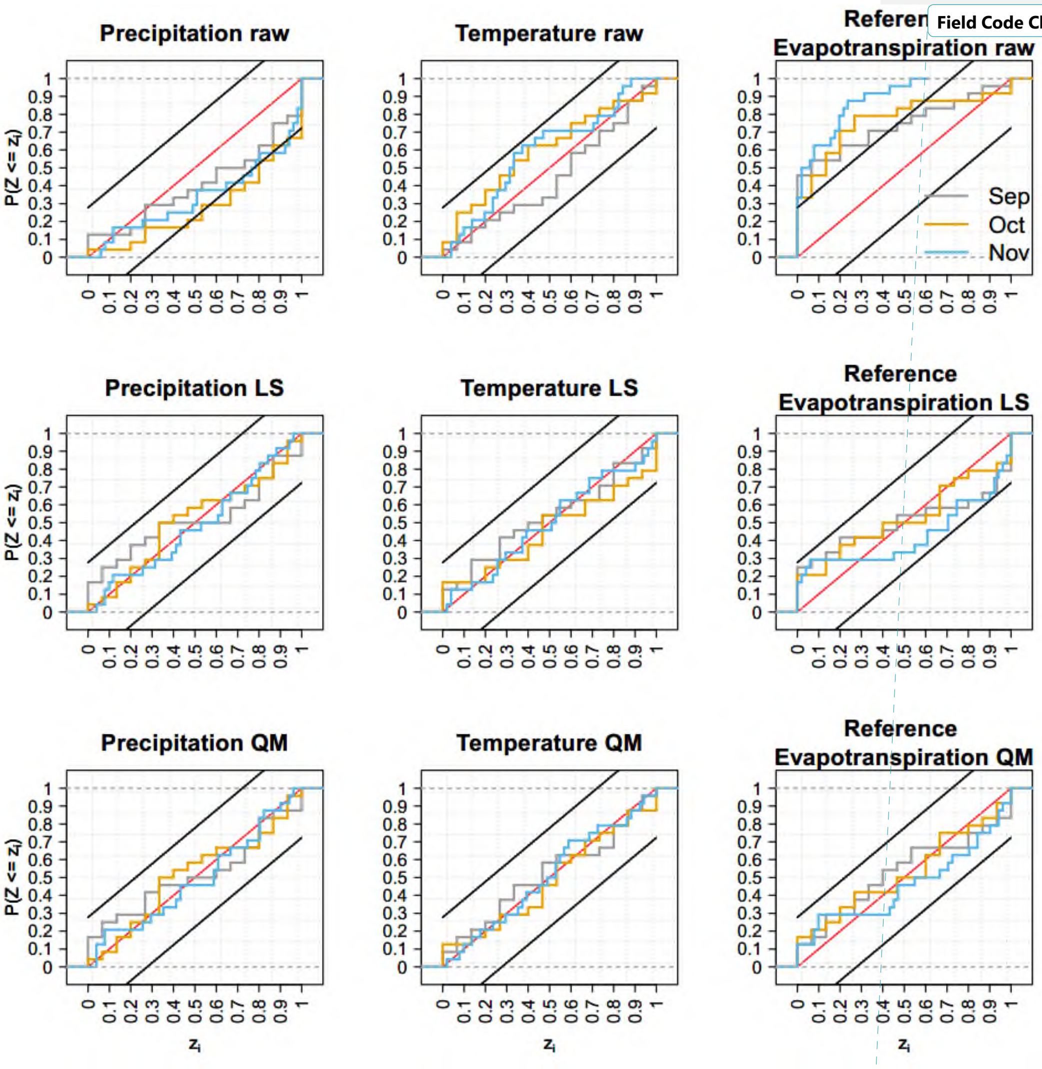
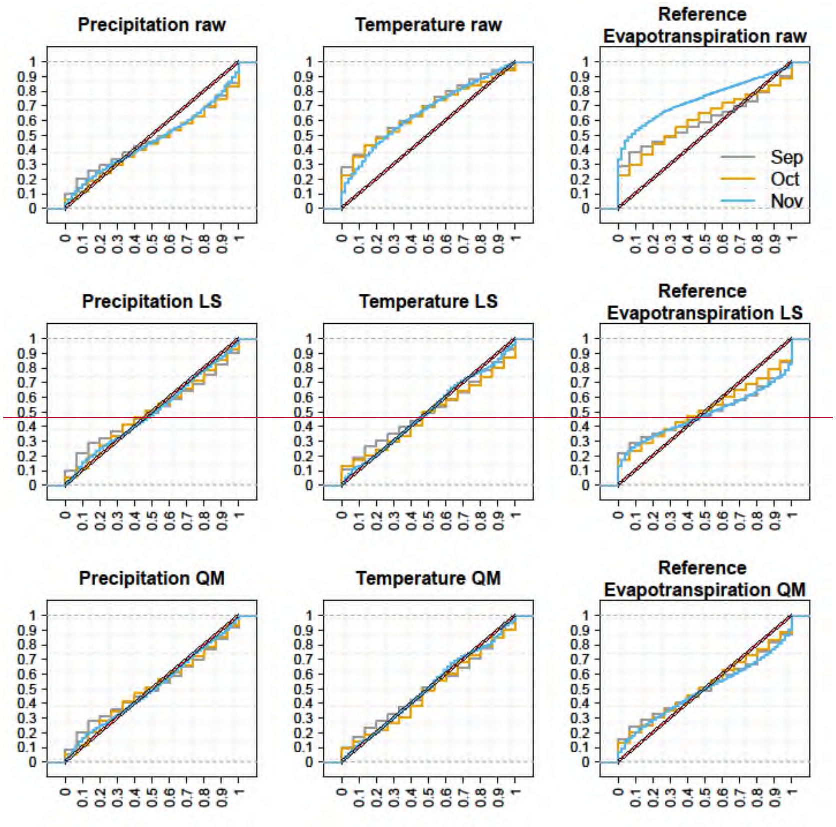


Figure S7: As in Fig. 7, but for spring.



Field Code Changed


Precipitation raw
Temperature raw
Reference Evapotranspiration raw
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Sep
Oct
Nov

Field Code Changed

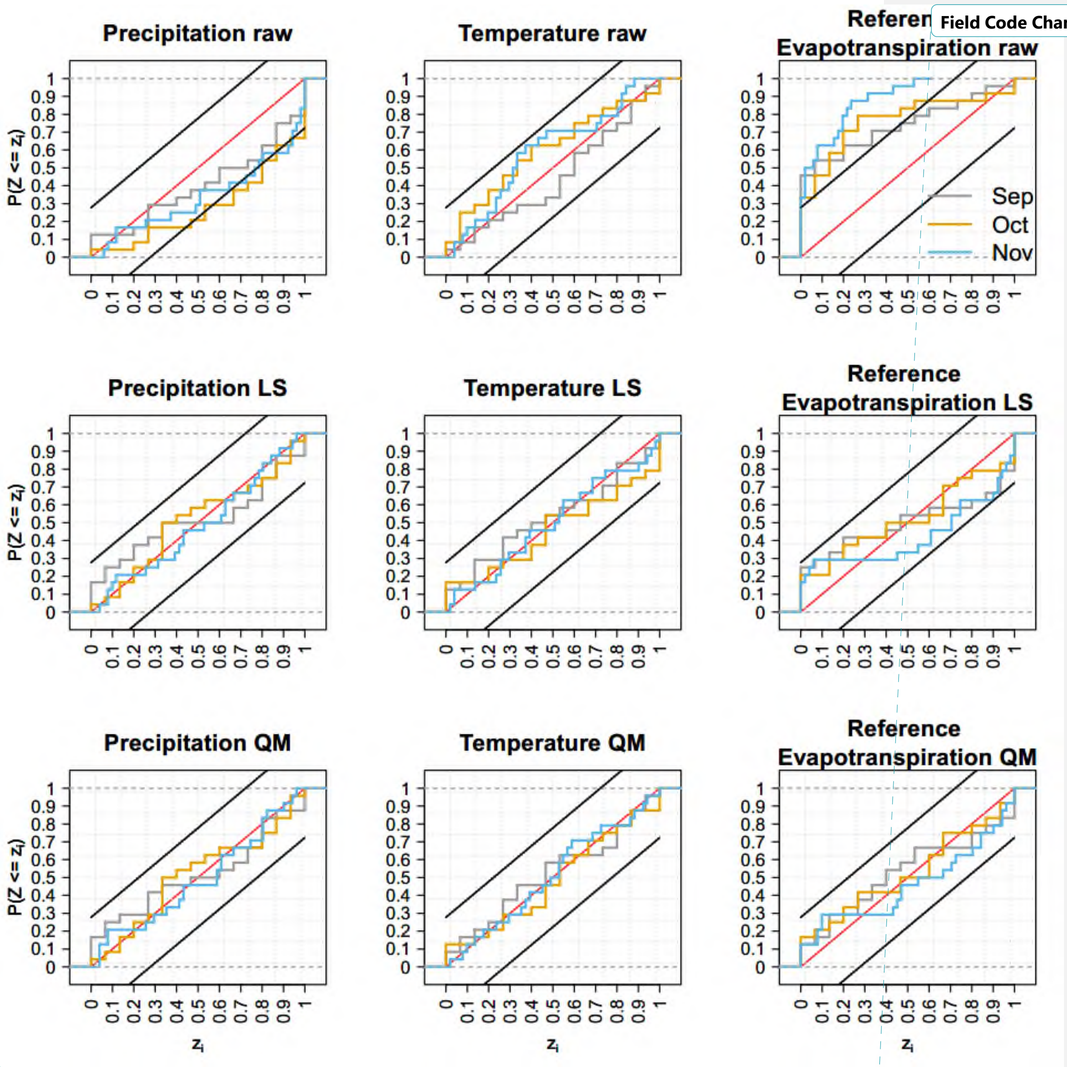
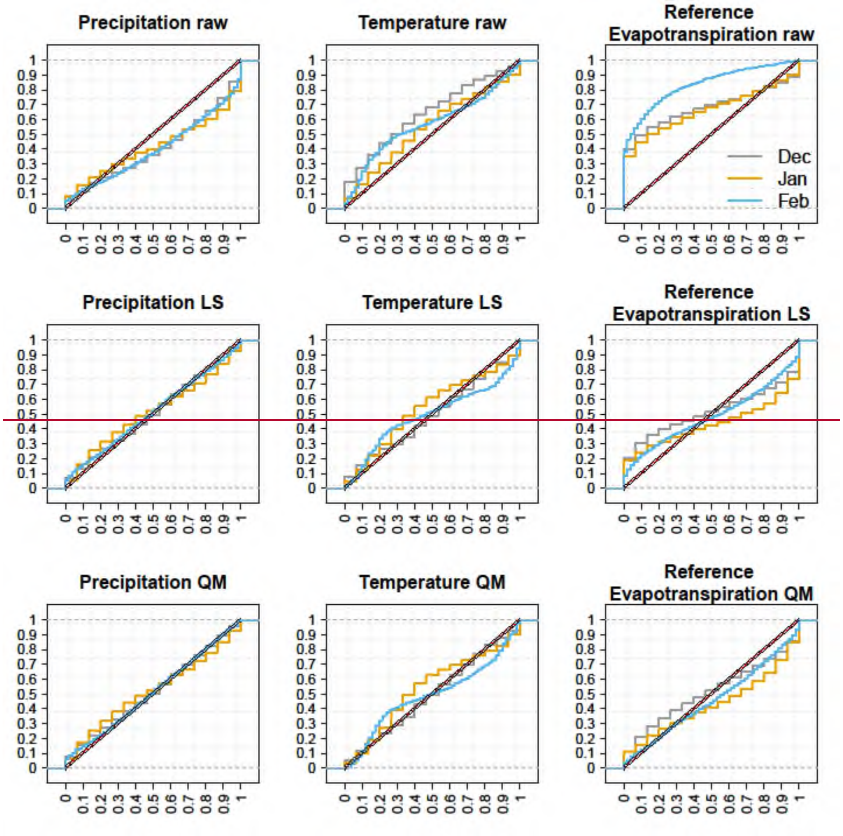
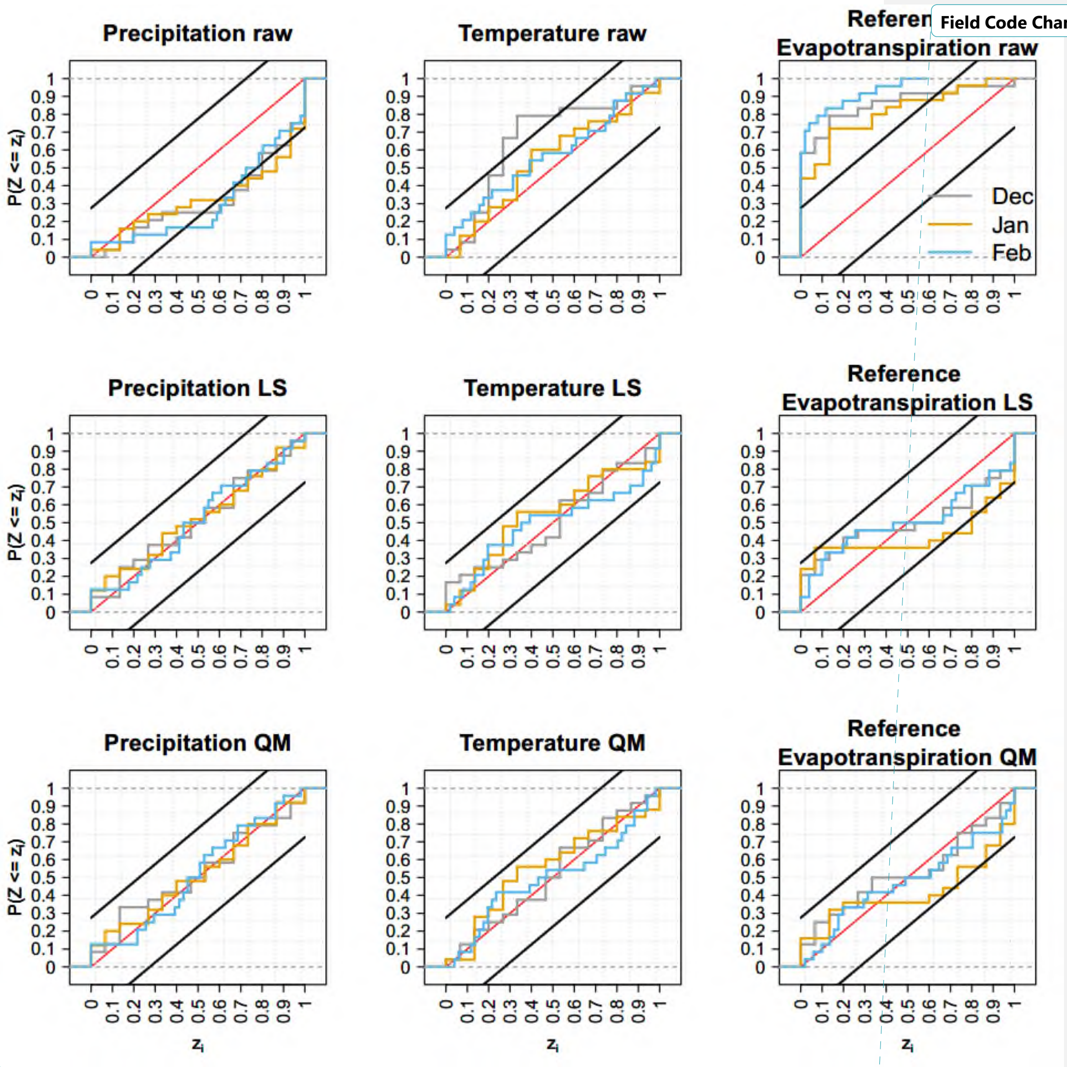


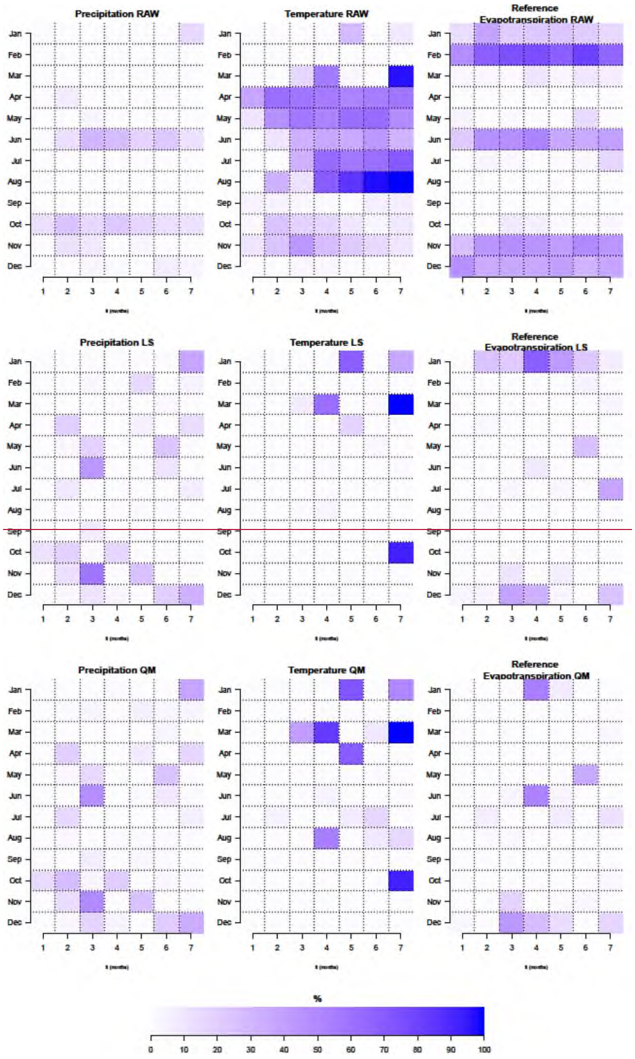
Figure S8: As in Fig. 7, but for autumn.

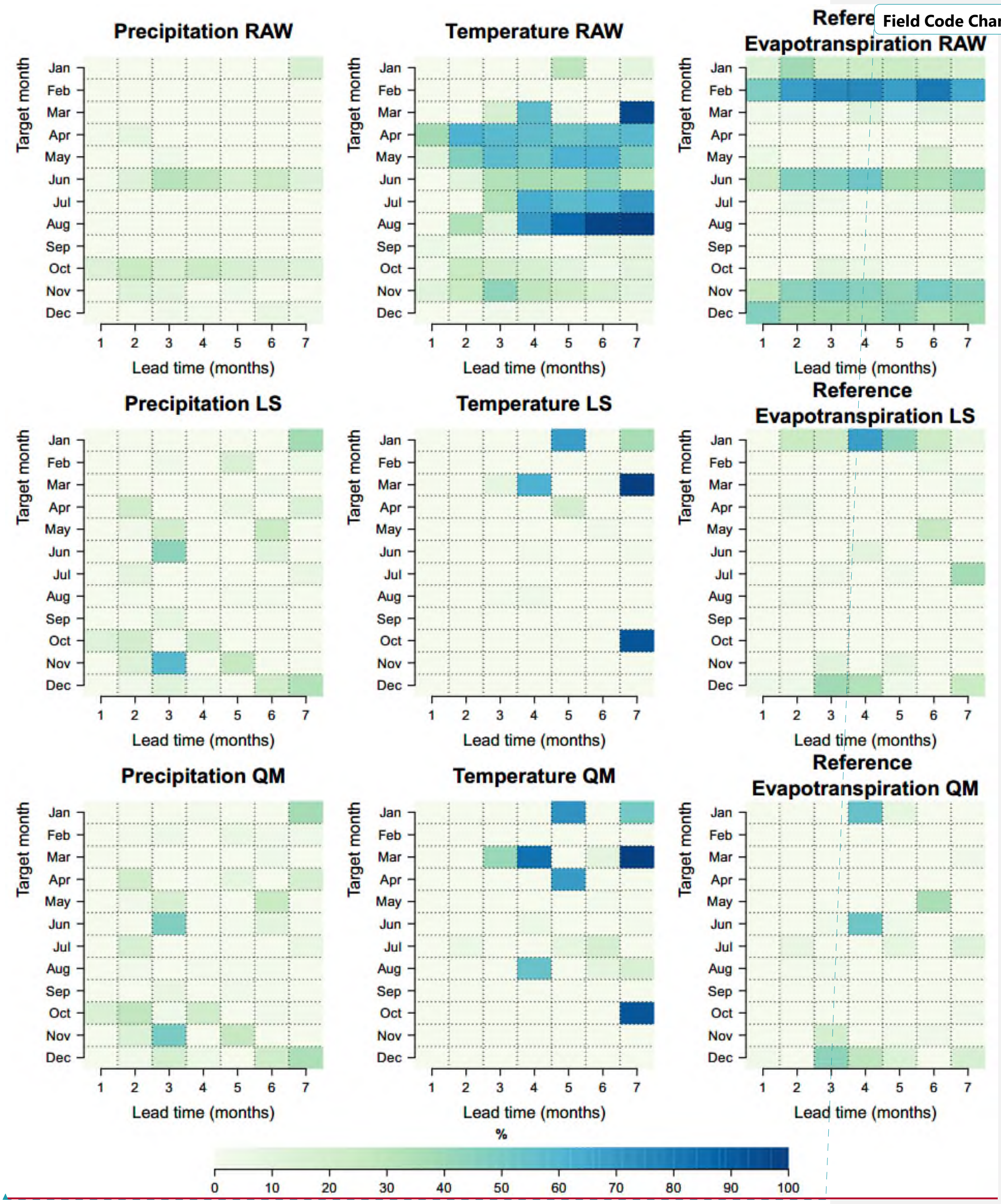

Precipitation raw
Temperature raw
Reference Evapotranspiration raw
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Dec
Jan
Feb
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1



Field Code Changed

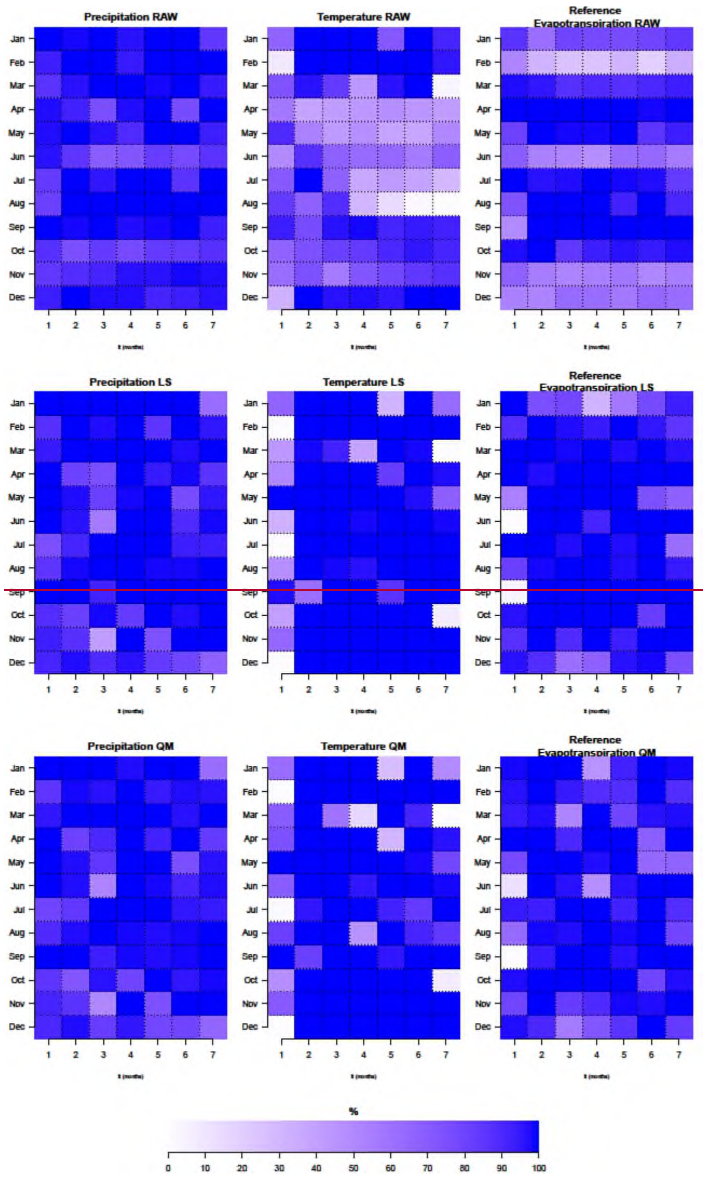
Figure S9: As in Fig. 7, but for winter.

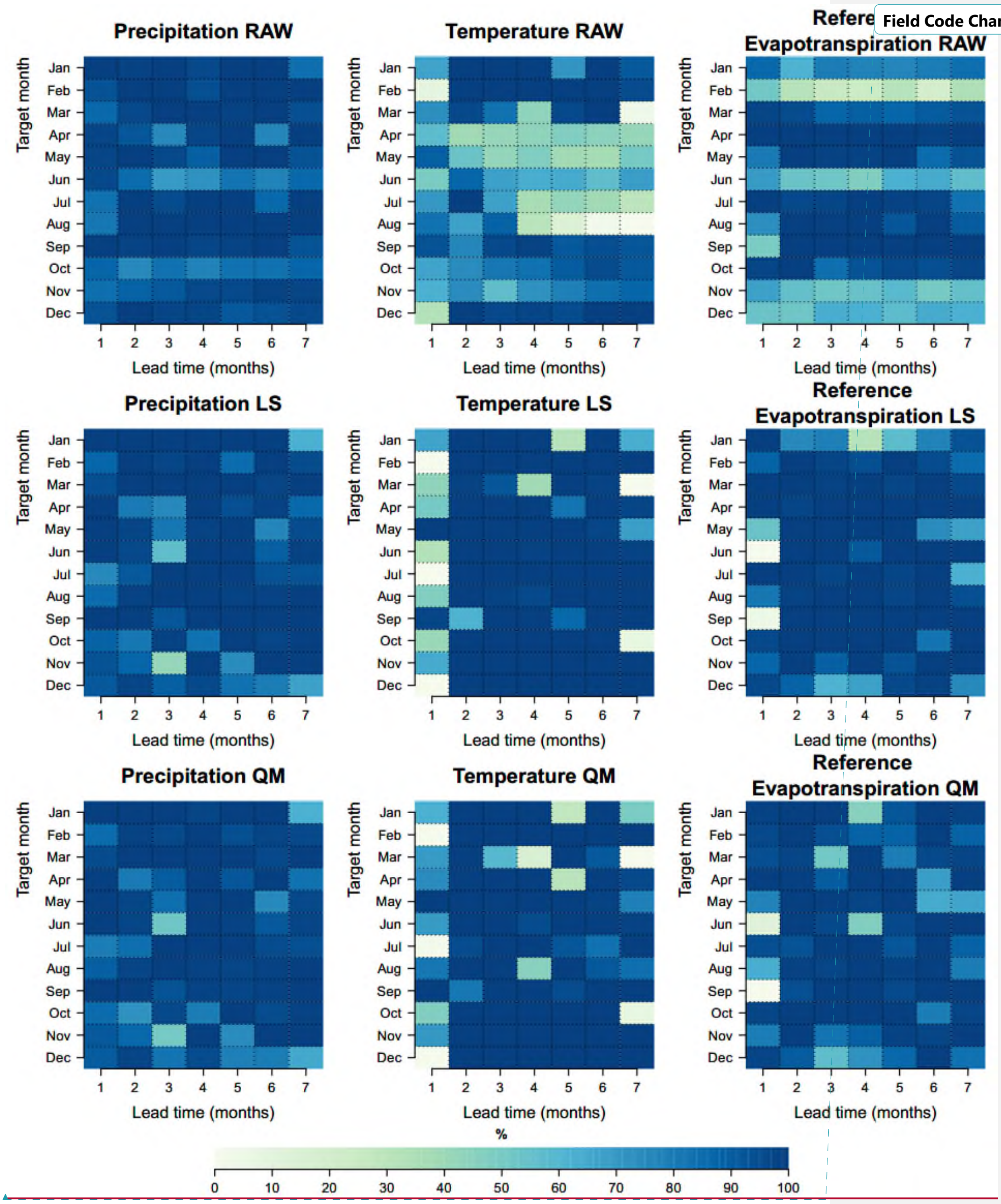

Precipitation RAW
Temperature RAW
Reference Evapotranspiration RAW
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
2
3
4
5
6
7
k (months)
%
0
10
20
30
40
50
60
70
80
90
100



Field Code Changed

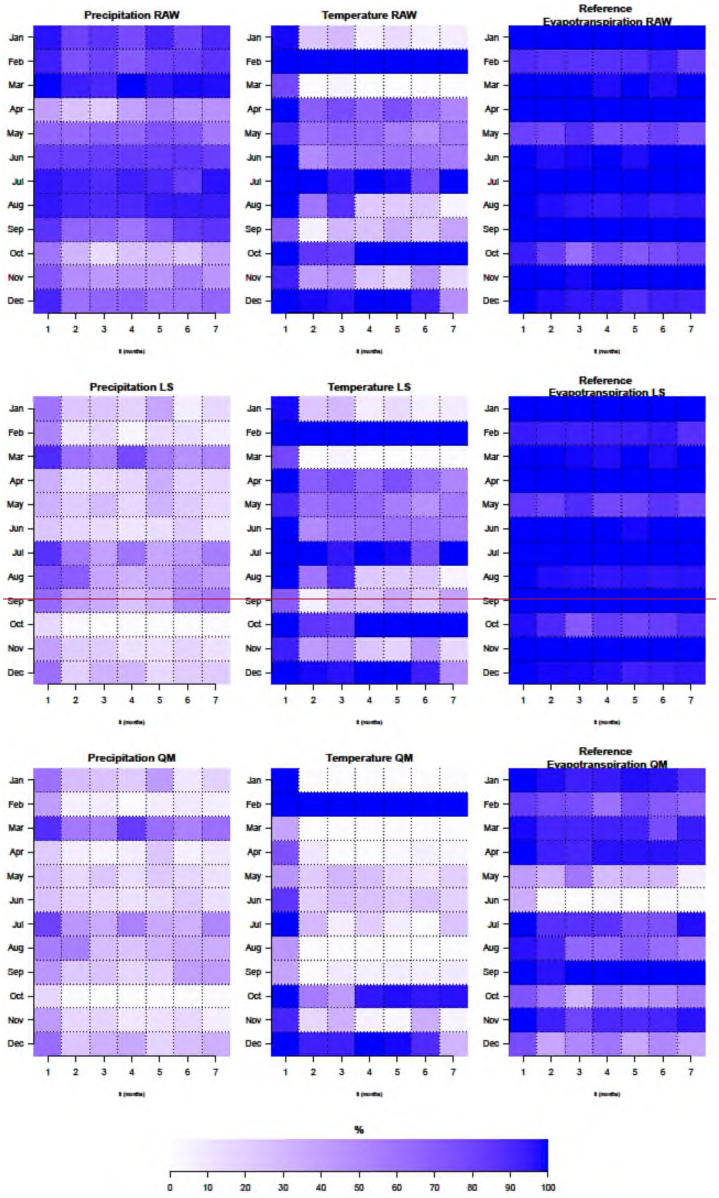
Figure S10: As in Fig. 9, but for negative accuracy (CRPSS in Eq. (8)).


Precipitation RAW
Temperature RAW
Reference Evapotranspiration RAW
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
2
3
4
5
6
7
t (months)
%
0
10
20
30
40
50
60
70
80
90
100

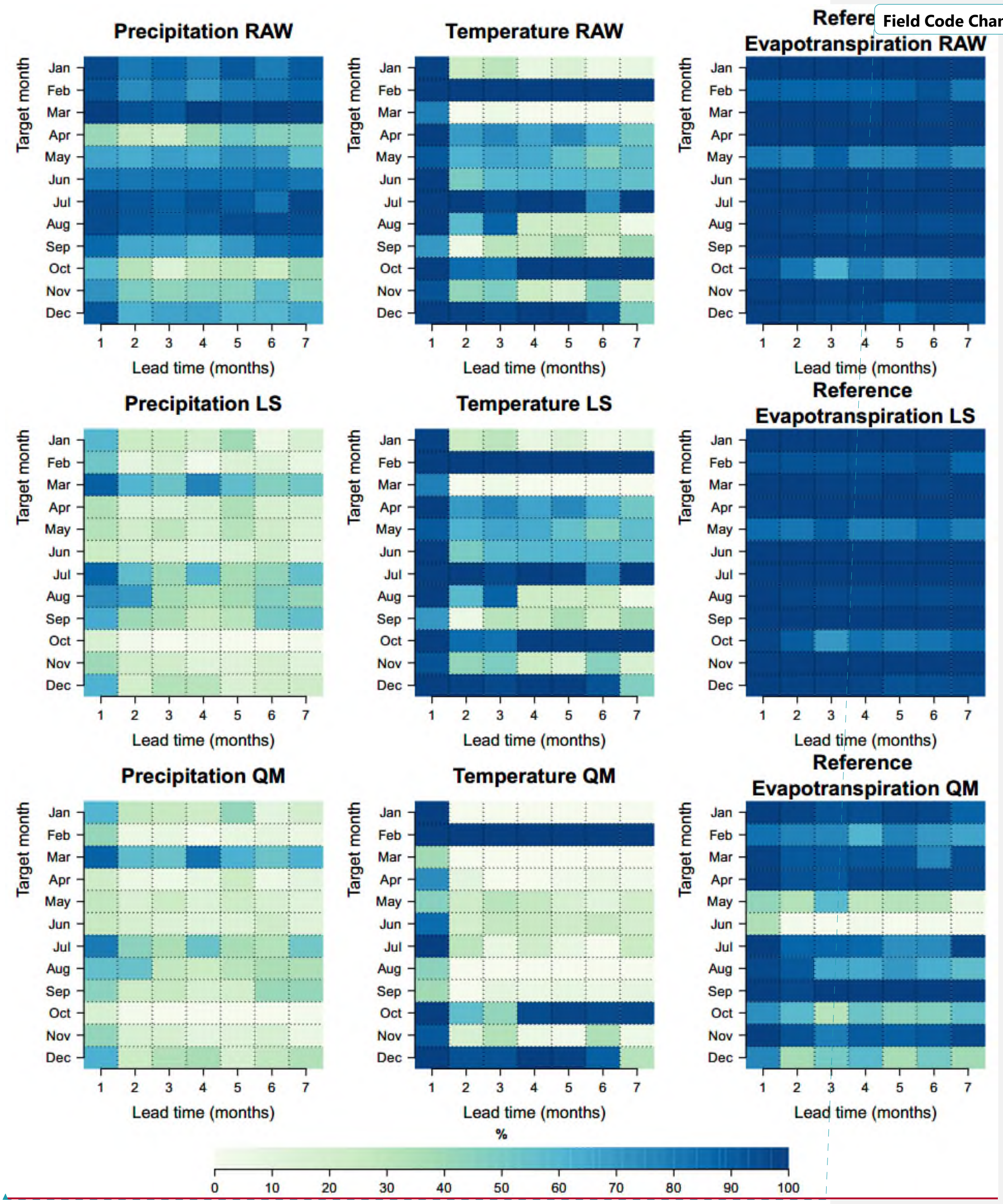


Field Code Changed

Figure S11: As in Fig. 9, but for equal accuracy (CRPSS in Eq. (8)).

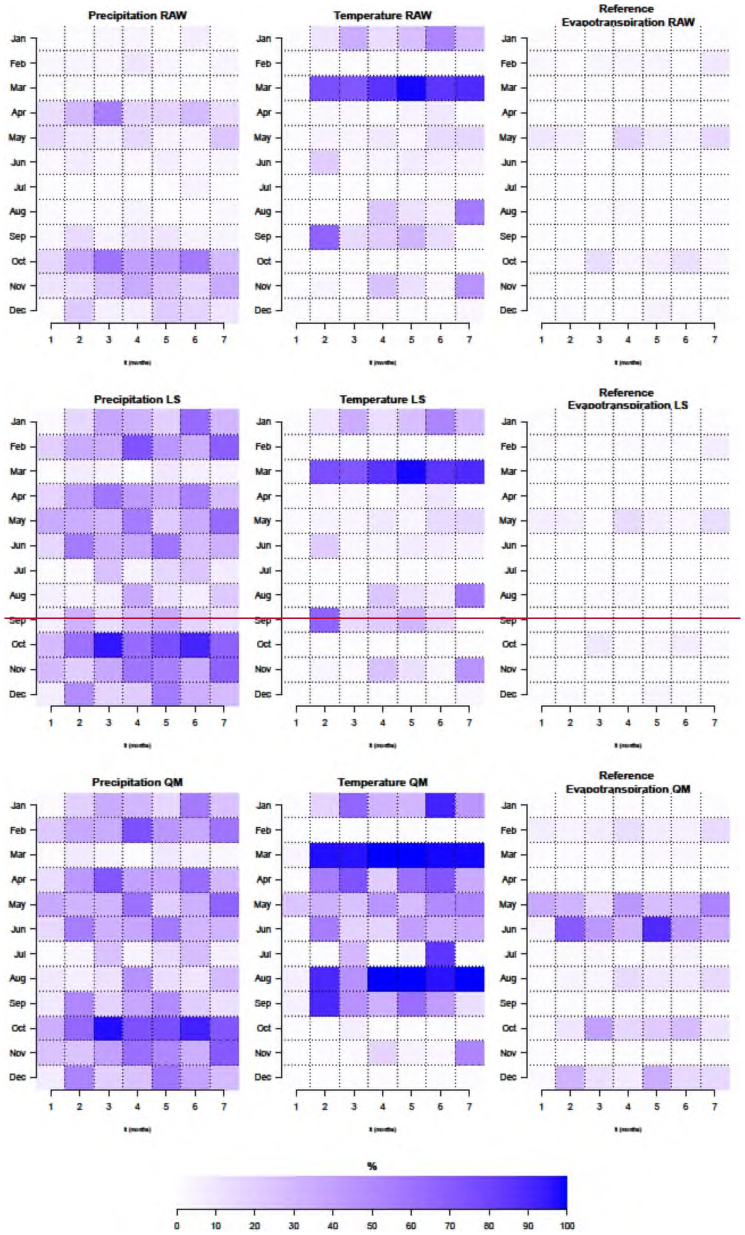


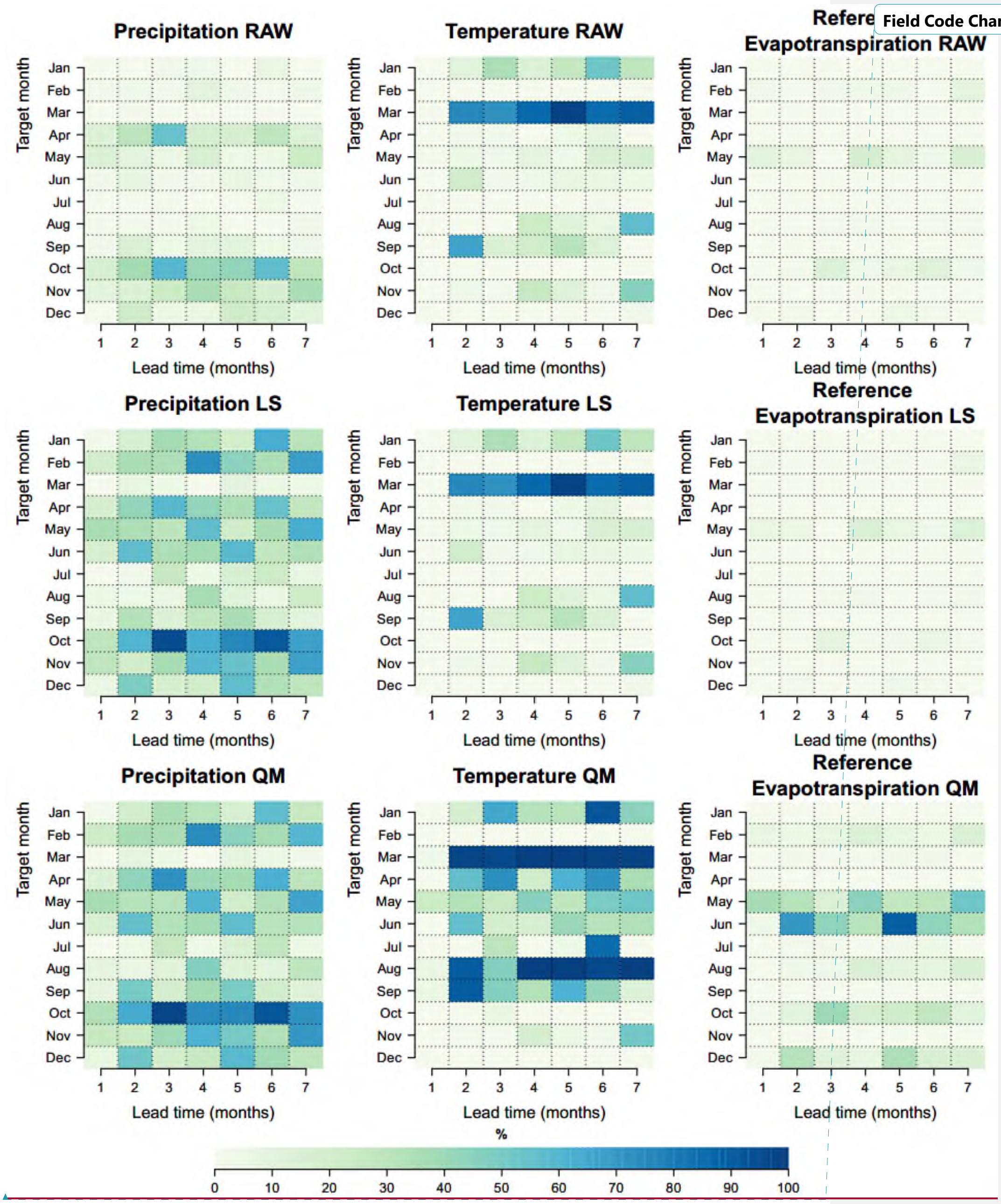
Precipitation RAW
Temperature RAW
Reference Evapotranspiration RAW
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
2
3
4
5
6
7
%
0
10
20
30
40
50
60
70
80
90
100



Field Code Changed

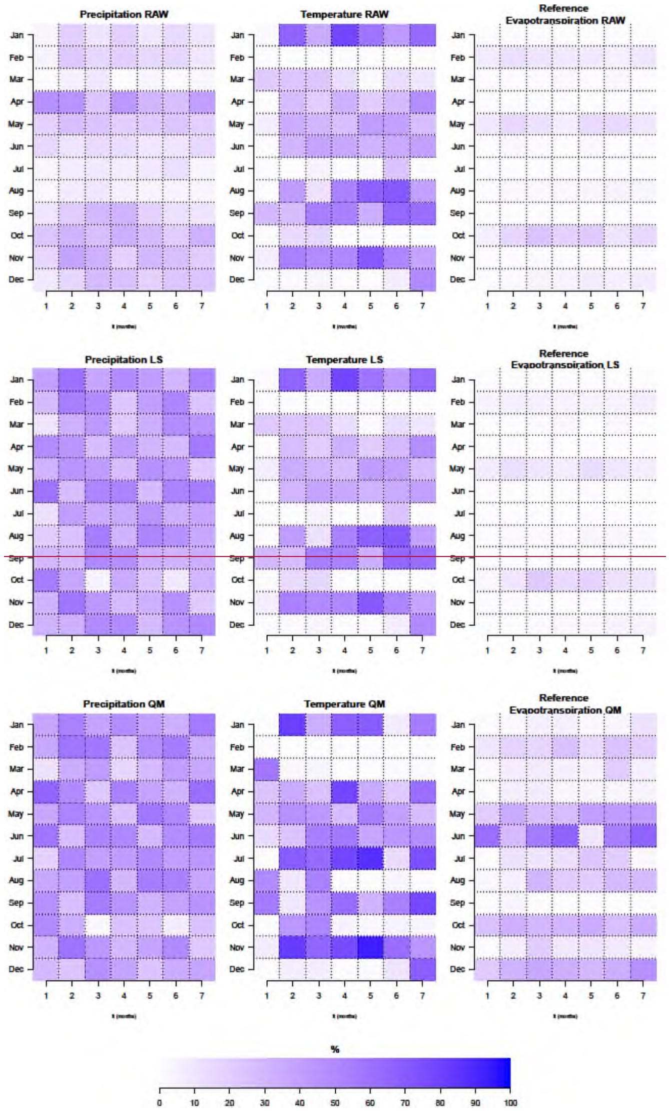
Figure S12: As in Fig. 9, but for positive sharpness (SS in Eq. (8)).

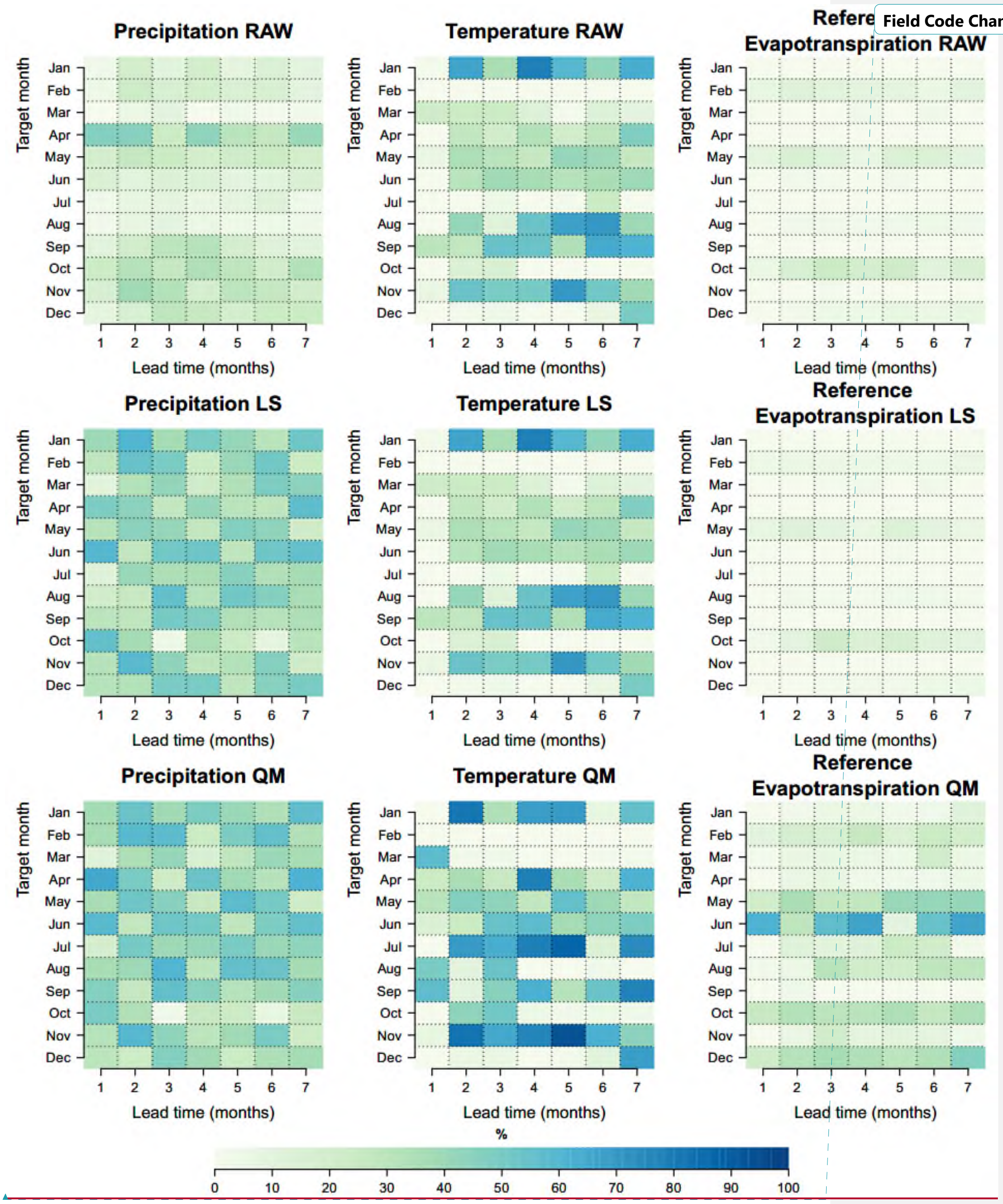

Precipitation RAW
Temperature RAW
Reference Evapotranspiration RAW
Precipitation LS
Temperature LS
Reference Evapotranspiration LS
Precipitation QM
Temperature QM
Reference Evapotranspiration QM
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
2
3
4
5
6
7
%
0
10
20
30
40
50
60
70
80
90
100



Field Code Changed

Figure S13: As in Fig. 9, but for negative sharpness (SS in Eq. (8)).





Field Code Changed

Figure S14: As in Fig. 9, but for equal sharpness (SS in Eq. (8)).