We highly appreciate the insightful comments that the reviewers have provided. Their comments and their suggestions to modifications will surely improve the quality of the paper.

Reviewer # 1

General comments:

I found the paper interesting to read, and it addresses some relevant scientific questions within the field of hydrology and seasonal forecasting. The methodology is clearly outlined, and the overall presentation is well structured.

Comment 1. However, I recommend that the introduction be expanded with more information about the use and skill of GCM based seasonal forecasts in the region, where I found some information to be lacking. Also, the main scientific conclusion needs more clarification in my opinion. Finally, I suggest some minor revisions, for which I refer to the specific comments below.

Thank you for bringing this issue up. We realize that connections to the companionship paper "Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J., and Jensen, K. H.: On the skill of raw and postprocessed ensemble seasonal meteorological forecasts in Denmark, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-366, in review, 2017" are lacking. We will make the connection more clearly in the updated manuscript version.

Comment 2. The language is generally precise and understandable, albeit a bit verbose sometimes. I recommend maybe trying to shorten some sentences, and to use the active tense more instead of the passive tense. Finally, see also the technical corrections below to improve the English.

We will improve the writing to ease the reading.

I do recommend the manuscript for publication, subject to minor revisions.

Specific comments

Comment 3. - p2, line 4: "perceived" : this gives the impression that the lack of reliability is a perception, do you mean a general observed fact here?

It is an observed fact (Weisheimer and Palmer, 2014). We will change the sentence accordingly to avoid confusion.

Comment 4. -p2, paragraph4, and further on in the manuscript: you say that seasonal forecasting in the region is a "challenging endeavor". I would like some more information here on the state of the art and skill of current seasonal predictions for temperature, precipitation, .. in the region. I realize that results are/will be presented in a different paper (Lucatero et al), but since this is still "under preparation", it hinders me a bit in comprehending the general skill of the seasonal meteorological forecasts you are using as input for your hydrological model. Could you provide some salient features at least? It would also aid in understanding the conclusions better. If there is no skill past a certain forecast range for example, it is important to appreciate this before using the forecast as input... Is there a published reference that could be useful here?

We will clarify the connection to the companion paper as stated before.

Comment 5. -p.3, paragraph3: concerning the DMI observations: I assume these are daily values? Please clarify. paragraph 4: does the hydrological model take snow melt into account? How important is snow in this area?

Yes, DMI observations are daily values, which will be clarified in the revised version of the manuscript. The model takes snowmelt into account by using a simple degree-day model formulation. Overall, snow processes are not important in the study area.

Comment 6. -p.3,section2.2: you use acronyms here that haven't been explained yet (LS and QM). And as mentioned above, I would prefer to have some of the most important features of the (as yet unavailable) Lucatero et al paper available here.

The comments are noted and they will be incorporated in the revised version.

Comment 7. - p.4, line 1: why the change in the number of ensemble members?

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.

Comment 8. - p.4, section 2.3: could you add an existing publication referencing to the Quantile Mapping method?

We will add a reference on Quantile Mapping (Zhao et al., 2017).

Comment 9. - p.5, concerning PIT diagrams. Is there a reason that you use these instead of e.g. verification rank (Talagrand) histograms? Since you are comparing ensemble members with observations (and not PDF's with observations), it seems less appropriate... Also, please clarify how you go from the ensemble to the CDF. Is it just the empirical CDF? Or do you use some smoothing, like fitting a gaussian to each ensemble (cfr. Grimit and Mass, 2004).

First, we make use of PIT diagrams for purely practical graphical reasons. We considered that it would be easier for the reader to visualize lines in Fig. 6 and Fig. 8 rather than bar plots in rank histograms. Moreover, we presume that if rank histograms were used instead, the conclusions would not change due to the connection between the rank histogram and the PIT diagram. In the updated version of the paper, we will make an attempt to demonstrate this or discuss the implications and appropriateness of our choice (PIT diagrams over rank histograms).

Related to the above and answering to your second question, we go from the ensemble to the CDF using empirical CDF. $z_i = P(X<=y_i)$ is the number of ensembles below or equal to the observed value divided by the number of ensembles (Page 5, lines 23-25). We will also clarify this in the new version of the manuscript.

Comment 10. - p.6, section 2.5. Low flow forecasting. You evaluate these in the same way as the monthly flow forecasts. Could you comment on timing errors? For example, how relevant would it be for water management if the forecast predicts the correct low flow, but in the wrong month? You also mention two other studies on low flows that exist. Is there a link with their results and yours? Do they use the same verification methodology?

Observed low flow is computed as the flow of the day with the minimum discharge over the seven month forecasting period (April-October). Forecasted low flow (for each ensemble) is computed in the same manner.

Timing errors will only be visible if forecasted low flow was chosen to be the discharge values of the day where low flow was observed, which is not the case here.

We will make the connection to the cited literature more clearly in the updated version of the manuscript.

Comment 11. - p.9, Summary and Conclusions. I am missing a bit a general conclusion that could be useful for end users of the hydrological forecasts. Are the current seasonal meteorological forecasts just not good enough? Could better postprocessing techniques ameliorate the situation? Data assimilation? The last paragraph seems very important, concerning catchment characteristics and taking advantage of hydrological memory.. but seems almost added as an afterthought. A few final sentences on the "best way forward" according to the authors could help here, and if a more explicit link to the results on seasonal forecasts of Lucatero et al could be made, this could provide a clearer overview in my opinion.

These comments are highly valid and we will improve the conclusions and discussion accordingly.

Technical Corrections

- p.1 line 11: forecasts (plural) line 37: "or outputs from, use..." : not grammatically correct, please rephrase - p.2 line 7: "should" instead of "may" ? line 25: "of ECMWF" instead of "from" line 31: "How do... compare to those..." (plural)

- p.4 line 2: should be "number of ensemble members"

- p.5 line 13: "Moreover..." -> this sentence needs to be rephrased line 15: should be "raw OR preprocessed" ?

- p.9 line 28: "effect on" instead of "in" line 33: "Thus, it seems..." instead of "It thus"

- p.10 line 13: "did not perform" instead of "did not performed" line 16: 'small' and 'large' errors (instead of 'high' and 'low') line 23: should be "..which in turn helps.." line 29: "particularly for" line 30: "which in turn translate"" please rephrase

Two additional comments:

- Figure 8 caption should explicitly mention "PIT diagrams"

- The reference "Molteni et al, 2011" does not include the journal (ECMWF Technical Memorandum 656)

All technical corrections will be incorporated in the revised manuscript.

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

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.

Reviewer # 2

Dear authors,

This paper is a nice well designed study on pre- and post-processed seasonal ensemble predictions. Such studies are needed to learn the limits and opportunities yielded by such prediction systems.

Comment 1. In some parts the referencing and discussion suffers from the fact that a companion paper dealing with the meteorological pre-processing is also in revision. The authors should better state that there is a companion paper and therefore some aspects concerning meteorology are not dealt here.

This issue was also brought up by Reviewer # 1. In the updated version of the manuscript we will improve the connection to the companion paper.

Comment 2. While it is nice to have a study dealing with both pre- and post-processing. The applied methodologies (mainly for the post-processing) are rather simple as compared to state-of-the-art methods. The authors should give more room on the discussion of their findings with respect to findings obtained by "higher-order" post-processing techniques.

We are aware that perhaps more sophisticated processing methods might have an impact on the quality of the forecasts. However, this comparison should be done in a consistent manner, which is a research question that is outside of the scope of this study. Our objective was to evaluate the benefits of pre- and post-processing using a fairly simple and popular method. We will address more clearly this issue on the Discussion section of the updated manuscript.

Comment 3. The issue of having in some cases 15 and in other cases 51 members should be addressed also in the results and discussion sections. I give in the commented manuscript a reference to have a look at.

We will carry out the evaluation as suggested . It will be addressed in the new version.

Comment 4. The discussion section should be extended and separated from the conclusions.

We will separate the discussion and conclusion sections in the revised manuscript and generally improve these two sections.

We respond to specific questions noted in the pdf of the manuscript in the following.

Comment 5 .Page 1. Line 21. Where do you see it in your outcomes?

We do not observe it here. However, we believe this to be an important component of the objective of the study (improving forecast quality) which can be dealt with in future research.

Comment 6 .Page 3. Line 3. How representative is this basin within Denmark and the Baltic region?

It represents a groundwater dominated catchment dominated by outwash coarse sandy materials representative of the western part of Denmark. Further it is located in a temperate climate and as such represents the Baltic region.

Comment 7. Page 3. How is estimated ET0 by DMI? How is estimated ET0 in the ECMWF products? How does this relate to the DMI product?

Both DMI observations and forecasts are estimated using the Makkink equation as explained in Page 3 Lines 39-40.

Comment 8. Page 4. Any reason for that? Wouldn't be easier to have 15 all the time? Can you put in all figure an asterisk on the months where you have 51 members?

As explained in the response of Comment 7 of Reviewer #1, the data we received from the reforecast provider has different ensemble members for the individual months. We will emphasize this in the revised manuscript and indicate in the figures when the analysis is based on 51 members.

Comment 9. Page 4. Do I understand right, all experiment start on January 1990 and run until the start of the forecast eg: 1.10.1993 or 1.4.2002 ... does the system not allow to save states?

The model is run once from January 1990 up until December 2014 saving initial conditions on the first day of each month.

Comment 10. Page 6. In the presented data for the year 2000 most issues on gauge 21 seems to be related to the cold season. Can you comment/confirm? The large bias could be also a sign of some sub-surface transfer of water in the headwater regions?

Yes, the bias is higher for the cold season. If there is some sub-surface transfer of water this should be accounted for in the model.

Comment 11. Page 6. The quality of raw forecasts (CRPSS) gradually decrease with lead time. This is easy to catch and agrees with our findings concerning monthly forecasts (See Fundel et al. paper cited before). Now, if I look at the CRPSS obtained after applying the pre-processors I cannot find any intuitive pattern in the skill of your system concerning lead time, while a pattern related to the season of initialization is intelligible using the bias metric. So there is good news here that the metrics improve, but at first glance one can think that this is not linked to processes. Can you comment?

We need to investigate this and will make comments on it in the revised manuscript.

Comment 12. Page 7 and 8. Can you please include in your reply and exemplary forecast with hydrograph and members of the different version you use? Also here it would be nice to see in the reply an example of post-processed forecast with the raw ensemble as a background.

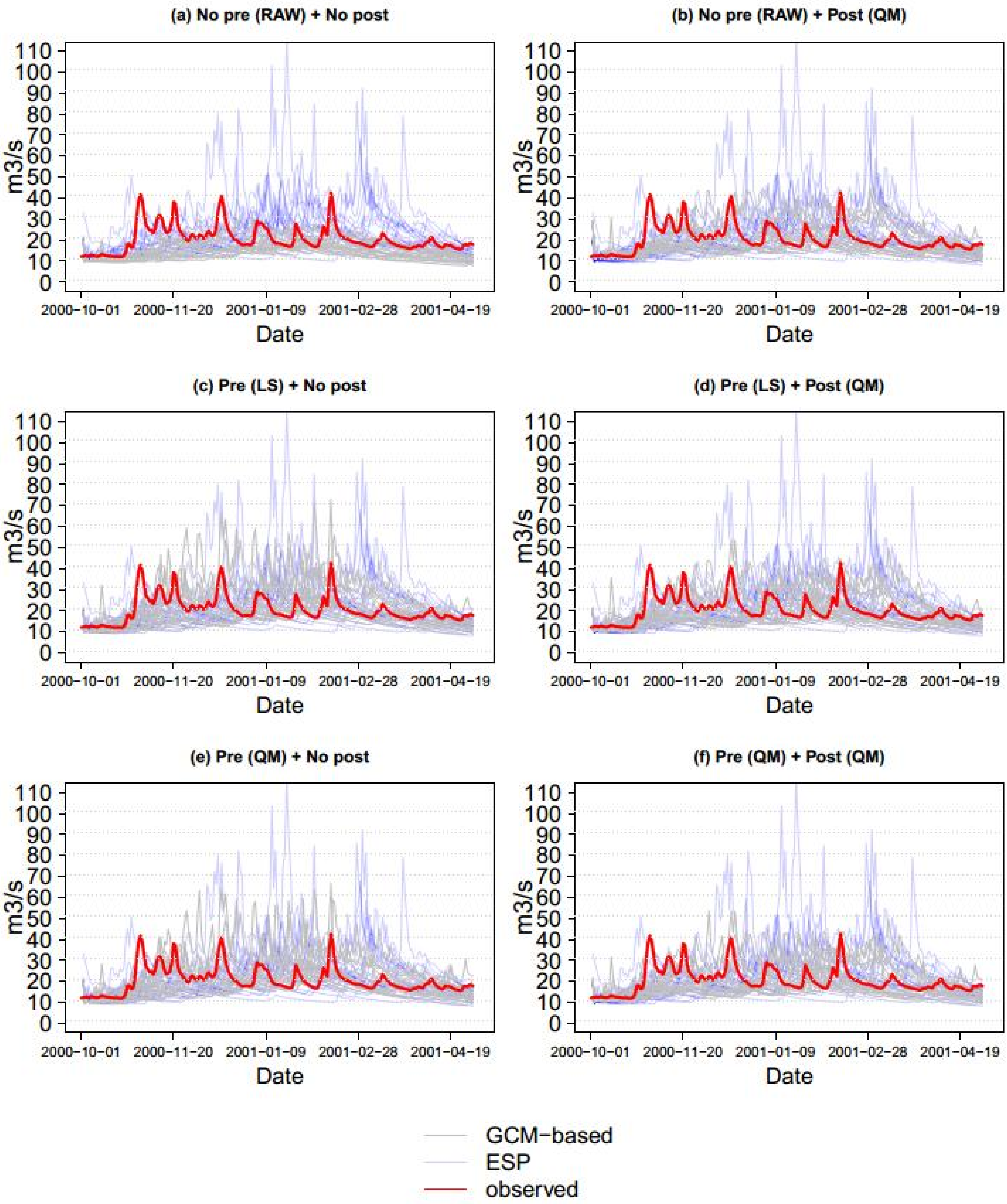


Figure 1: Example of a forecast of daily streamflow (initialized 01-10-2000, outlet station 82) generated with different strategies. Pre and Post stand for preprocessing and postprocessing, respectively. RAW, LS, QM and ESP stand for raw forecast, linear scaling , quantile mapping and Ensemble Streamflow Prediction respectively.

Comment 13. Page 7. Would be of course nice to have also a figure for one month, and different lead times, same applies to Figure 8

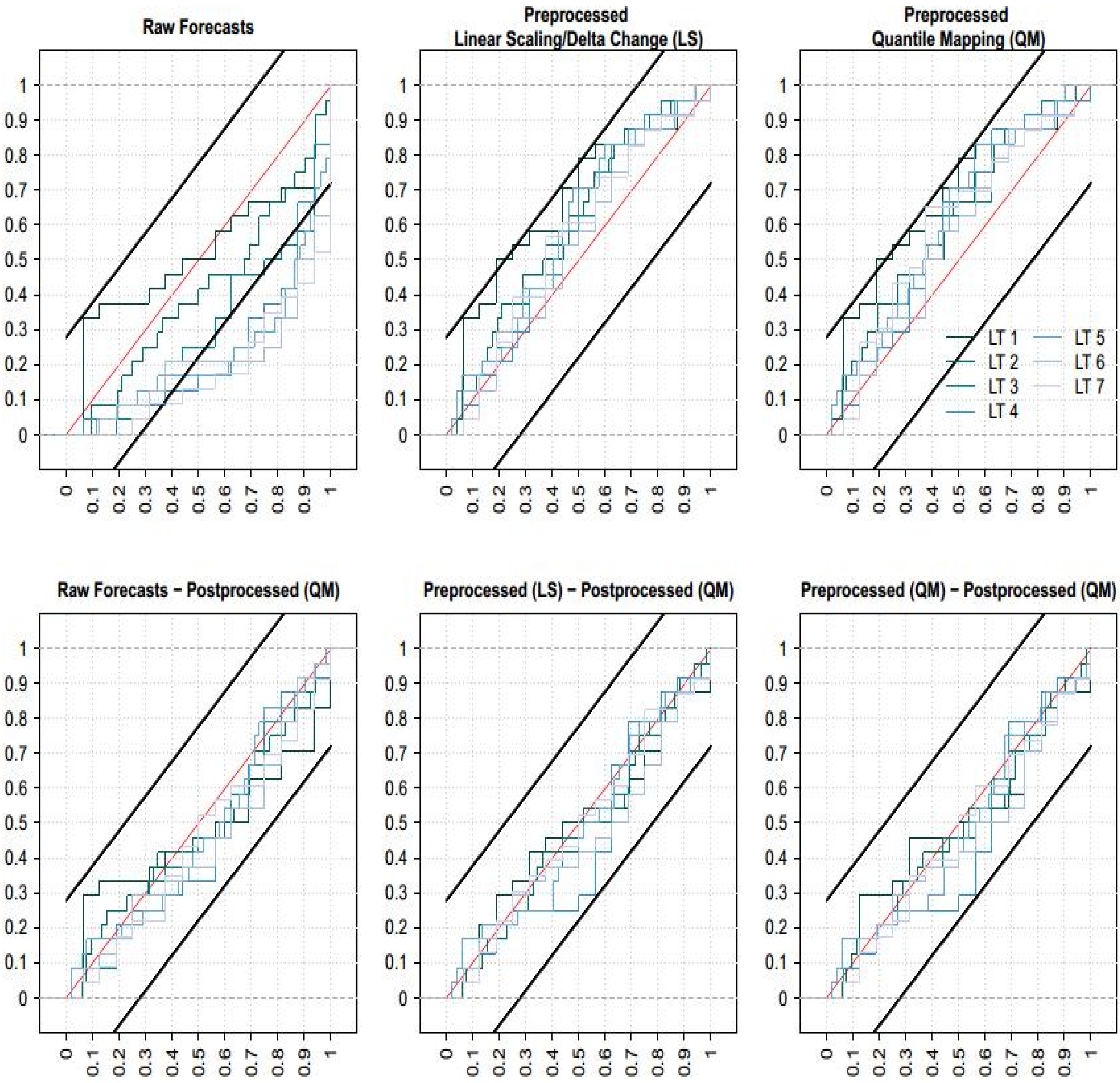


Figure 2: PIT diagrams of target month December at different lead times (LT) for different forecast strategies. For example, LT 3 corresponds to the December forecast initialized in October.

Comment 14. Page 8. Page 12. This is the companion paper in HESSD, isn't?

Yes, this is the one. As stated above, we will make the connection clearer on the revised version of the manuscript.

Comment 15. Page 16. As a curiosity, have you tried to Postprocess ESP?

No, this was not the purpose of the study.

Comment 16. Page 20. Can be that February works bests, because you have 51 members in the raw forecasts? How would this look like if you pick up 15 random members several times (bootstrapping) and then you average it?

We believe it has more to do with compensational errors during winter than the difference in number of ensemble members as discussed in Sect. 3.5. However, in order to remove the effect of varying ensemble size, we will make the analysis as you suggest and add the resulting changes (if there are any) in the revised version of the manuscript.

List of relevant changes

1. The connection to the companion paper "Lucatero, D., Madsen , H., Refsgaard, J.C., Kidmose, J., and Jensen, K. H.: On the skill of raw and postprocessed ensemble seasonal meteorological forecasts in Denmark, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10. 5194/hess-2017-366, 2017." was made clearer in Introduction and Section 3.5.

2. An additional experiment to address the issue with varying ensemble size was done. Results are explained in Sect. 3.2. Minor modifications to Fig. 5, Fig. 7 and Fig. 10 were done to specify months/lead times with 51 ensemble members.

3. Comment 11 of Page 6 of Reviewer #2 was clarified in Section 3.5.

4. Sentences were shortened. Changes can be seen on the marked-up file.

5. Section 4. Summary and conclusions of the original submission were separated to create and extend a Discussion section.

6. Appendix 5 and Fig. A1 were removed as a similar figure is included in " Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 1267–1277, 2007.". The reader is then referred to this paper.

Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: effect of preprocessing and postprocessing on skill and statistical consistency

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. In the present study we analyze the effect of bias adjustments in both meteorological and streamflow forecasts on skill and ~~reliability~~ statistical consistency of monthly ~~average~~ streamflow and low flow forecasts. Both raw and pre-processed meteorological seasonal forecasts from the European Center for Medium-Range Weather Forecasts (ECMWF) are used as inputs to a spatially distributed, coupled surface – subsurface hydrological model based on the MIKE SHE code. Streamflow predictions are then ~~in order to generate streamflow predictions~~generated up to seven months in advance. In addition to this, we postprocess streamflow predictions using an empirical quantile mapping technique ~~that adjusts the predictive distribution in order to match the observed one~~. Bias, skill and statistical consistency are the qualities evaluated throughout the forecast generating strategies and we analyze where the different strategies fall short to improve them. ECMWF System 4-based streamflow forecasts tend to show a lower accuracy level than those generated with an ensemble of historical observations, a method commonly known as Ensemble Streamflow Prediction (ESP). This is particularly true at longer lead times, for the dry season and for streamflow stations that exhibit low hydrological model errors. Biases in the mean are better removed by postprocessing that in turn is reflected in the higher level of statistical consistency. However, in general, the reduction of these biases is not enough to ensure a higher level of accuracy than the ESP forecasts. This is true for both monthly mean and minimum yearly streamflow forecasts. We ~~highlight~~ discuss the importance of including a better estimation of the initial state of the catchment, which ~~will~~ may increase the capability of the system to forecast streamflow at longer leads.

1 Introduction

Seasonal streamflow forecasting encompasses a variety of methods that range from purely data based to entirely model based or hybrid methods that exploit the benefits of each (Mendoza et al., 2017). Data driven methods find empirical relationships between streamflow and a variety of predictors. These relationships are then used ~~and use these~~ to derive forecasts for the upcoming seasons. Different predictors can be used depending on the relative importance they have on the regional hydroclimatic conditions. Predictors that have been used include large scale climate indicators such as el Niñ~~n~~io or the North Atlantic Oscillation, (Schepen et al., 2016; Shamir, 2017; Wang et al., 2009; Olsson et al., 2016), precipitation and land temperature (Córdoba-Machado et al., 2016), state of the catchment in the form of streamflow, soil moisture, groundwater storages or snow storages that can be derived either by the use of a hydrological model, therefore the term 'hybrid' (Robertson et al., 2013; Rosenberg et al., 2011), or by means of observed antecedent conditions (Robertson and Wang, 2012).

Model based systems include a hydrological model in the forecasting chain. Differences between forecasting frameworks may arise in the forcings, the initialization framework and/or the hydrological model structure and parameters. Focusing on

Formatted: Font: (Default) Times New Roman, 10 pt, Not Bold, English (United States) the forcing, one can either use observed meteorology from previous years, a method that is commonly known as Ensemble Streamflow Prediction (ESP) (Day, 1985), or outputs from~~, use~~ General Circulation Models (GCM) ~~outputs~~ (Crochemore et al., 2016; Wood et al., 2002, 2005; Wood and Lettenmaier, 2006; Yuan et al., 2011, 2013, 2015, 2016). In principle, the latter should be more suitable in providing skilful forecasts as GCMs are able to capture the evolving chaotic behavior of the atmosphere, whereas the ESP approach assumes that what has been observed in the past can be used as a proxy for what will happen in the future, an assumption that requires stationary climate conditions. On the other hand, the ~~perceived~~ lack of reliability GCMs have in forecasting atmospheric patterns at long lead times preclude ~~them for~~their use in weather impacted sectors (Bruno Soares and Dessai, 2016; Weisheimer and Palmer, 2014). For example, a previous study on the skill of ECMWF System 4 in Denmark, concluded that, in general, precipitation forecast bias in the catchment area was in general around -25% (Lucatero, et al., 2017). This bias, together with the sharpness of forecasts, lead to a mild positive skill limited to the first month lead time (Lucatero, et al., 2017). These results are in accordance to skill studies with focus on a similar area (Chrochemore, et al., 2017).

This is the reason why pre- and postprocessing ~~may~~ should be performed when using GCM forecasts to force a hydrological model to eliminate biases intrinsic in climate and hydrological models. In the context of this study, preprocessing refers to any method that improves the forcings, i.e. precipitation and temperature, used in the hydrological forecasting system. Postprocessing refers to the improvements done to the outputs of the hydrological model, e.g. streamflow. In this respect, postprocessing also corrects errors in hydrological models that cannot be eliminated through calibration (Shi et al., 2008; Yuan et al., 2015; Yuan and Wood, 2012).

A couple of studies have quantified the effects on streamflow skill by preprocessing either seasonal (Crochemore et al., 2016) or medium range ~~forecasts~~ (Verkade et al., 2013) forecasts. Other studies have assessed the efficiency of postprocessing streamflow forecasts only (Bogner et al., 2016; Madadgar et al., 2014; Ye et al., 2015; Zhao et al., 2011;Wood and Schaake, 2008). To our knowledge, only Roulin and Vannitsem, (2015); Yuan and Wood, (2012) and Zalachori et al., (2012) have compared the additional gain in skill of doing both pre- and postprocessing. The previous studies have shown that improvements made by preprocessing the forcings do not necessarily translate into improvements in streamflow forecasts (Verkade et al., 2013; Zalachori et al., 2012). Improvements are larger when postprocessing is done, and a combination of pre- and postprocessing provides the best results (Yuan and Wood, 2012; Zalachori et al., 2012). To our knowledge, only Yuan and Wood, (2012) have made this evaluation in the context of seasonal forecasting.

The present study focuses on the following aspects: (i) the evaluation of the use of a GCM to generate seasonal forecasts, (ii) the study of the effect that pre- and postprocessing have on streamflow forecasts 1-7 months ahead, and (iii) the effect of hydrological model biases in forecast skill evaluations. This is done by a combination of the following methodological choices. First, we make use of seasonal meteorological forecasts ~~from~~ of ECMWF System 4 (Molteni, et al., 2011). Secondly, the hydrological simulations ~~are based on~~use an integrated physically-based and spatially distributed model based on the MIKE SHE code (Graham and Butts, 2005). Thirdly, our evaluation focuses on three forecast qualities: bias, skill and reliability~~, with s~~. Skill is measured using ESP as a reference and focusing on both accuracy and sharpness. Finally, the focus here is to evaluate forecasts of monthly average streamflow throughout the year and low flows during the summer. The catchment serving as basis of our study is~~ of a~~ groundwater dominated ~~catchment~~ located in a region where seasonal forecasting is a challenging endeavor (Lucatero, et al., 2017). The following questions are then addressed:

(1) How does GCM generated forecasts compare to those of the ESP approach?

Formatted: Font: Not Italic

(2) What is the effect of pre- and postprocessing on streamflow forecasts in terms of bias, skill and statistical consistency. And more specifically, is there one single approach, or a combination of several that reduces the bias and augments skill and statistical consistency?

(3) What is the effect that hydrological model bias has on the evaluation of pre- and postprocessed streamflow forecasts?

2 Data and Methods

The following sections give a description of the methodology followed in this study. A graphical depiction of the steps carried out can be seen in Fig. 1.

2.1 Area of study, observational data and hydrological model

The present study is carried out for the Ahlergaarde catchment located in West-Jutland, Denmark (Fig. 2)~~. The catchment covers~~with ~~an area~~size of 1044 km^2. ~~and is l~~Located in one of the most irrigated zones in Denmark with 55% of the area covered with agricultural crops such as barley, grass, wheat, maize and potatoes. The remaining area is distributed as follows: grass (30%), forest (7%), heath (5%), urban (2%), and other (1%) (Jensen and Illangasekare, 2011).

The climatology of the area is shown in Fig. 3. Climate in the Ahlegaarde region is mainly driven by its proximity to the sea towards the west. The mean annual precipitation, reference evapotranspiration and discharge for the period 1990-2013 is 983 mm, 540 mm and 500 mm, respectively. The hydrology of the catchment is groundwater dominated due to the high permeability of the top geological layer, which consists mainly of sand and gravel. ~~The 1990-2013 average monthly streamflow has a maximum in January (60 mm) and a minimum during the summer (25 mm).~~ Another consequence of the geological composition of the surface layer is that overland flow rarely happens. A more detailed description of the geology of the area can be found in Kidmose et al., (*unpublished work*).

Daily ~~Precipitation~~ precipitation (P), temperature (T) and reference evapotranspiration (E_{T0}) data are retrieved from the Danish Meteorological Institute (DMI; Scharling and Kern-Hansen, 2012). The dataset spatial domain~~, which~~ covers Denmark with a 10 km grid resolution for P and a 20 km resolution for T and E_{T0}. ~~(Scharling and Kern-Hansen, 2012) with~~ P is corrected for systematic under catch due to wind effects (Stisen et al., 2011, 2012) and ET_0 is derived using the Makkink formulation (Hendriks, 2010). Finally, daily ~~Streamflow~~ streamflow observations are ~~taken~~ retrieved from the Danish Hydrological Observatory (HOBE) (Jensen and Illangasekare, 2011) datasets.

Formatted: Font: Not Italic, Subscript

The hydrological simulations for this study are ~~based~~ grounded on a physically-based, spatially distributed, coupled surface-subsurface model that simulates the main hydrological processes such as evapotranspiration, overland flow, unsaturated, saturated and stream flows and their interactions. The model is based on the MIKE SHE code (Graham and Butts, 2005). Groundwater flow is described by the governing equation for three-dimensional groundwater flow based on Darcy's law. Drain flow is considered when the groundwater table exceeds a drain level. Surface water flow in streams is simulated by a one-dimensional channel flow model based on kinematic routing, while a two dimensional diffusive wave approximation of the St. Venant equations is used for overland flow routing. Finally, a two layer approach is used for the simulations of unsaturated flow and evapotranspiration (Graham and Butts, 2005). Snow is not an important process on the study area, therefore, the model takes snowmelt into account by using a simple degree-day model formulation. The horizontal numerical discretization is 200 meters, whereas the vertical discretization is based on six numerical layers whose dimension depends on the geological stratigraphy. Model parameters were calibrated against groundwater head and discharge using an automated optimizer, PEST (Parameter Estimation) version 11.8 (Doherty, 2016) for the 2006-2009 period. Parameters to be calibrated were selected based on a sensitivity analysis study. These are: hydraulic conductivities for ten geological units, specific yield, specific storage, drain time constant, detention storage, river-groundwater conductance and root depth of ten vegetation types. The reader is referred to Zhang et al., (2016) and Kidmose et al., (*unpublished work*) for further details on the calibration procedure.

2.2 Forecast generation: GCM-based and ESP

As seen in Fig. 1, P, T and E_{T0} forecasts are taken from the ECMWF System 4 (RAW), preprocessed ECMWF System 4 (Linear Scaling; LS and Quantile Mapping; QM), and historical observations (ESP). The European Center for Medium-Range Weather Forecasts (ECMWF) offers a seasonal forecasting product that currently is in its version number 4 (Molteni et al., 2011). An attempt to reduce the biases intrinsic in ECMWF System 4 led to what we refer to as preprocessed forecasts. The reader is referred to Lucatero, et al., (~~*in preparation*~~,2017) for details of the evaluation of both ECMWF System 4 and preprocessed forecasts for Denmark. The spatial resolution of the raw forecasts is 0.7 degrees in the latitude and longitude. Forecast ~~that~~ were interpolated to a 10 km grid to match the resolution of the observed ~~precipitation~~ grid. For the Ahlergaarde catchment, a forecast-observation data for the 1990-2013 period is extracted ~~total of 24 grid points were extracted that cover~~for 24 grid points covering the study area, leading to a sample size of 24 years. ~~for the 1990-2013 period, leading to a sample size of 24 years~~. Finally, E_{T0} is computed using the Makkink formulation (Hendriks, 2010) that takes as inputs T and incoming shortwave solar radiation from the ECMWF System 4 forecasts.

Daily ~~Raw~~ raw and preprocessed forecasts are initialized on the first day of each calendar month ~~and have a lead time of 7 months with a daily temporal resolution~~with a 7 month lead time. The number of ~~ensembles~~ ensemble members varies with month, 15 for January, March, April, June, July, September, October and December, and 51 for the remaining months. The number of ensembles is higher for February, May, August and November to aid in improving forecasts for the most predictable seasons. ESP forcings are taken from the observation record with each year acting as an ensemble member. ~~The v~~Values are taken from the start of each calendar month, with a 7-month lead time in order to match the lead time of the ECMWF System 4 forecasts. ~~The~~ Since the year to be forecasted is withdrawn from the ensemble,~~. Thus,~~ the number of ensemble members for the ESP is 23. Both the ECMWF System 4 generated forecasts and ESP share the same hydrological initial conditions for forecasts initiated on the same month. These are computed from a spin-up run starting in January 1990 and up until 2013. Initial states are saved on the first day of each calendar month~~the time of the forecast initialization~~. Forecasts are then run on a daily basis up to seven months.

2.3 Preprocessor and Postprocessor

Preprocessed forcings for the hydrological model were retrieved from data of the companion paper Lucatero, et al., (~~*in preparation*~~ 2017). The authors used two well-known bias correction techniques, ~~namely, Linear Scaling/Delta Change (hereafter, LS) and Quantile Mapping (QM)~~LS and QM. In LS the ensemble is adjusted with a scaling factor, either by multiplication (for P and E_{T0}) or addition (T). The scaling factor is computed as the ratio or difference between the averages of the ensemble mean and the observed mean for a specific month, lead time and location, with the sole purpose of adjusting the mean.

QM (Zhao et al., 2017) matches the quantiles of the ensemble distribution with the quantiles of the observed distribution in the following way:

$$f^{*}_{k,i} = G^{-1}\left(F\left(f_{k,i}\right)\right) \qquad (1)$$

Formatted: Font: Not Italic

Formatted: Font: Not Italic

where G and F represent the observed and the ensemble distribution functions, respectively, for forecast-observation pair i, for $i = 1, \cdots, M$ with M being the number of forecast-observation pairs. $f_{k,i}$ represents ensemble member k, $k = 1, \ldots, N$ where N is the ensemble size, and $f^*_{k,i}$ represents the corrected ensemble member k. F is an empirical distribution function trained with all ensemble members at a given month for a given lead time and location. *G* and *F* are fitted on a leave-one-out-cross-validation mode, i. e., forecast and observation pair *j* are withdrawn from the sample. For example, for a forecast of target month April initialized in February, F is computed using all ensemble members, comprising 30 (days) times 23 (number of years in the training sample minus the year to be corrected) times the ensemble size of that particular month (15 or 51). The same is done for G. Linear extrapolation is applied to approximate the values between the bins of F and G and to map ensemble values and quantiles that are outside the training sample.

QM is the only method used for postprocessing in the present study as no striking differences in both bias and skill were found between LS and QM in Lucatero, et al., (2017~~*in preparation*~~). Moreover, QM shows more satisfactory results for the correction of forecasts in the lower tail of the distribution and for correcting forecasts that also exhibit underdispersivity (Lucatero, et al., ~~(~~2017~~in *preparation*)~~).

Formatted: Font: Not Italic, English (United States)

Formatted: English (United States)

Formatted: Font: Not Italic

Formatted: Font: Not Italic

2.4 Performance metrics

~~To evaluate first the raw forecasts~~ The performance of raw, pre- and postprocessed forecasts is evaluated. Our main focus is the following four qualities: ~~and the improvement after preprocessing and postprocessing, we check for four qualities:~~ bias, skill in regards to accuracy and sharpness, and statistical consistency. Bias is the measure of under or overestimation of the mean of the ensemble in comparison with the observed values (Yapo et al., 1996):

$$PBias = \left(\frac{\sum_{i=1}^{M} \bar{f}_i}{\sum_{i=1}^{M} y_i} - 1 \right) \cdot 100 \tag{2}$$

where $\bar{f}_i$ and y_i represent, respectively, the ensemble mean and the observed values for forecast-observation pair i of a particular month, lead time and location. If the value in Eq. (2) is negative, we have an underprediction, and conversely an overprediction if the value is positive.

Secondly, we compute the continuous rank probability score CRPS (Hersbach, 2000) as a general measure of the accuracy of the forecasts. The computation of the score is as follows

$$CRPS = \frac{1}{M} \sum_{i=1}^{M} \int_{-\infty}^{\infty} \left[P_i(x) - H(x - y_i) \right]^2 dx, \tag{3}$$

where $P_i(x)$ represents the CDF of the ensemble for forecast-observation pair i, $H(x - y_i)$ is the Heaviside function that takes the value 0 when $x < y_i$ or 1 otherwise. y_i is the verifying observation of forecast-observation pair *i* ~~*i* of *M* forecast-observation pairs~~. Sharpness for forecast-observation pair i is measured as the difference between the 25% and the 75% percentiles. The average of these differences along the forecast-observation record is then used as a measure of sharpness. Both the CRPS and ~~the~~ sharpness scores are then given in the units of the variable of interest, i.e. m^3/s for streamflow. ~~Moreover, both scores of a system that is perfect is zero~~Both scores are positive oriented, i.e., the lower the value, the more accurate or sharper is a forecast. A skill score can be then computed in the following manner

$$Skill = 1 - \frac{Score_{sys}}{Score_{ref}} \tag{4}$$

where, for the present study, $Score_{sys}$ is the score of streamflow forecasts generated either with raw, or preprocessed ECMWF System 4, or ~~the~~ postprocessed forecasts, $Score_{ref}$ is the score value of our reference system, the ESP. The range of the skill score in Eq. (4) is from $-\infty$ to 1, and values closer to 1 are preferred. Negative values indicate that, on average, our system does not manage to beat the ESP. Hereafter, we denote the skill with respect to accuracy as CRPSS and the skill in terms of sharpness as SS. In order to evaluate the statistical significance of the differences of skill between GCM generated forecasts and ESP, we use a two-sided Wilcoxon-Mann-Whitney test (WMW-test) at the 5% significance level (see Hollander et. al., 2014).

Since the number of ensemble members varies from month to month, the value of the skill scores for months with larger ensemble size will be more favorable. Although the purpose of the present study is not to make an in depth analysis of the effect of changing ensemble size, we utilized a bootstrapping technique to make the reader aware of the possible gains in skill due to increased ensemble size. This is accomplished by computing the skill scores of a random selection of 15 of the 51 ensemble members for February, May, August and November as in Jaun et al., (2008). This step is performed ~~a thousand~~1000 times ~~collecting the values of the skill score~~. The final value of the skill score of interest is then the average of ~~the thousand saved values~~these.

Finally, in order to evaluate the statistical consistency between predictive and observed distribution functions we use the Probability Integral Transform (PIT) diagram. The PIT diagram is the cumulative distribution function (CDF) of $z_i = P(X \le y_i)$, where z_i is the value of the cumulative distribution function that the observed value attains within the ensemble distribution for each forecast-observation pair *i*. Note that the PIT diagram is the continuous equivalent of the rank histograms (Friederichs and Thorarinsdottir, 2012) and its mainly used to evaluate statistical consistency of a continuous predictive CDF. However, in this study, the z_i's are based on the empirical CDF of the ensemble members at a given lead time. Note that the evaluation of the appropriateness of the choice of PIT diagrams over rank histograms for ensemble forecasts is out of the scope of the present study. For a forecasting system to be statistically consistent, meaning that the observations can be seen as a draw of the predictive CDF, the CDF of the z_i's should be close to the CDF of a uniform distribution in the [0,1] range. Deviations from the uniform distribution signif~~yies~~ bias in the ensemble mean and spread (see Laio and Tamea, 2007) ~~as explained in Appendix 1~~. Finally, ~~In~~ in order to make the test for uniformity formal, we make use of the Kolmogorov confidence bands. The bands are two straight lines, parallel to the 1:1 diagonal and at a distance $q(\alpha)/\sqrt{M}$ where $q(\alpha)$ is a coefficient that depends on the significance level of the test, i.e., $q(\alpha = 0.05) = 1.358$ (see Laio and Tamea, 2007; D'Agostino and Stephens, 1986) and ~~N~~ M is again the number of forecast-observation pairs. The test for uniformity is passed if the CDF of the z_i's lies within these bands.

2.5 Low flow forecasting

Low flow forecasts can be used for optimizing groundwater extractions for irrigation. The years for which the predicted low flows are above the prescribed minimum can be exploited and utilized for crops with a higher irrigation demand that may increase economic returns. Here we focus on forecasts initiated in April. For the purposes of this study, low flows are defined as the flow of the day with the minimum yearly discharge (m^3/s) that usually happens during July to September (Fig. 3). Note that timing errors are not an issue here due to the computation choice~~s~~ of low flows. Observed low flow is computed

Formatted: Font: Not Italic, Subscript as the flow of the day with the minimum discharge over the seven month forecasting period (April-October). Forecasted low flow (for each ensemble) is computed in the same manner. Timing errors will only be visible if forecasted low flow was chosen to be the discharge values of the day where low flow was observed, which is not the case here.~~Low flow forecasts are evaluated using the same skill scores as for monthly flow forecasts~~. Studies that have focused their attention to situations of low flow or hydrological drought in the context of seasonal forecasting exist~~:~~ (Fundel, et al., 2013, Demirel et al., ~~(~~2015~~)~~; Trambauer et al., ~~(~~2015)~~)~~ documenting the possibility to extract skillful forecasts months ahead for low flow/drought scenarios. Finally, low flow forecasts are evaluated using the same skill scores as for monthly flow forecasts, i. e., using ESP as a reference forecast. This will allow for an evaluation of the added value of GCM-based low flow forecasts more rigorous than the aforementioned studies provide.

3 Results

3.1 Hydrological model evaluation

Figure 4 shows the results for simulated streamflow at the upstream station 21 and the downstream station 82. The focus of the evaluation is done for daily values during the period from 2000-2003. As a preliminary evaluation, we computed the percent bias (PBias) and the Nash-Sutcliffe model efficiency coefficient (NSE) for the complete observed-simulated record (1990-2013). There is, in general, a good agreement in timing between observed and simulated values. The visual inspection of the hydrographs reveal, however, an amplitude error that is more pronounced at the upstream station 21, especially during the winter season. Evidence for this is also reflected by the high values of bias and the negative NSE for this station (NSE = -0.85). Furthermore, a scatter plot of simulated and observed low flows for the 24 years shows an overestimation of the low flows that is more pronounced at the upstream station (Fig. 4). At the outlet station 82 there is a better behavior in terms of bias and NSE, with an overestimation of only 1.7% and a NSE of 0.73. Moreover, for this station there is a better agreement in both the high and low flows through the year. The latter can be verified by looking at the scatter plot of the low flows (Fig. 4), with the majority of points lying close to the 1:1 diagonal.

Due to the poor performance in the upstream station 21, in the following sections (3.2 - 3.4) we will discuss the skill and consistency of the different approaches for forecast improvement with a focus on the outlet station only. The large biases in the upstream station, combined with the structural biases of the meteorological forecasts seem to inflate the skill of the streamflow forecasts. This will be further discussed in Section 3.5.

3.2 Streamflow forecasts forced with raw meteorological forecasts

The bias and skill of the monthly streamflow forecasts forced with raw ECMWF forecasts are shown in the first row of Fig. 5. The X-axis represents the different lead times in months, while the Y-axis represents the target month. For example, the bias of November with lead time 2 represents the value of bias for a forecast in that month initiated on October 1. This bias is in the [-30,-20%] range. In general, the absolute bias increases with lead time, and usually moves from an overprediction (or mild underprediction) to a large negative bias at longer lead times.

Figure 5 also shows the skill of accuracy and sharpness. The months with statistically significant differences in skill between the ESP and ECMWF System 4 forecasts are represented with a black circle. There is a connection between bias and skill of accuracy in the sense that months with a higher bias tend to be the ones with lower or non-existent skill. The opposite also holds, months with milder bias tend to be the months where the forecast is improving over the reference forecast to a higher degree. This is by no means surprising, as the CRPS penalizes forecasts that have biases.

The CRPSS is negative, except for some months during winter and at short lead times for which a forecast generated with raw ECMWF System 4 forcings improves accuracy up to 40% compared to ESP. As for the case of bias, skill depends on lead time, reaching its most negative values for forecasts generated 7 months in advance. One important feature is the high skill that a forecast generated with ECMWF raw forcings has in terms of sharpness. Figure 5 shows that this quality is present in the majority of target months and lead times. Note, however, that sharpness is only a desirable property when biases are low. In this case, the width of the raw forecasts is smaller than that of the ESP, indicating overconfidence, when biases are high.

The results of the bootstrapping procedure for the computation of the skill score due to accuracy indicate that, by reducing the ensemble size to 15, there is a reduction of skill as expected. However, this reduction does not change the main conclusion. Months with positive skill due to accuracy remain, in general, positive. For example, ~~skill score~~CRPSS of February streamflow forecasts at lead time 1 is 0.31 for 51 ensemble members. After the bootstrapping experiment with the reduction to 15 ensemble members, the skill score is mildly reduced to 0.29. In order to make the reader aware of the possible increase of skill due to increased ensemble size, green ~~stars~~crosses in Fig. 5 (and ~~and~~ subsequent figures dealing with skill due to accuracy), ~~will~~ represent the target months and lead times with 51 ensemble members.

Statistical consistency of the raw forecasts is visualized on the first column of the PIT diagrams in Fig. 6 for winter and summer (first and second row, respectively) at lead time 1. Kolmogorov confidence bands are also plotted for a graphical test of uniformity at the $\alpha = 0.05$ level. For the sake of brevity, the remaining seasons and lead times are not shown. For the particular seasons and lead time shown, statistical consistency seems to be achieved only for the wettest months (Dec-Feb). The explanation for this particular behavior will be given in Sect. 3.5. Early spring and November forecasts are also able to pass the uniformity test (not shown). Summer forecasts together with late spring and autumn months (May, September and October) show a significant underprediction, which prevent them to pass the uniformity test. Statistical consistency appears to get worse as lead time increases, in accordance with the deterioration of the bias in Fig. 5.

3.3 Streamflow forecasts forced with preprocessed meteorological forecasts

The second and third rows of Fig. 5 show the bias and skill of streamflow forecasts generated with preprocessed forcings from ECMWF System 4 using the LS and the QM method, respectively.

Several conclusions can be drawn when comparing forecasts using the preprocessed and raw forcings. First, ~~the~~ biases are clearly improved, especially for longer lead times. For example, for October forecasts from lead time 3 to 7 months, biases are reduced from the [-40,-30%] to the [-20,10%] range for LS and to the [-15,20%] range for QM. There are, however, no obvious differences between the two preprocessing methods, which seem to perform equally well in reducing biases. Secondly, three features on accuracy are seen. The first one is that, also for accuracy, there are no obvious differences in skill between the two preprocessing methods. Furthermore, there seems to be a reduction of skill for the winter months and March at the first month lead time. These months are the only ones with a statistically significant skill using the raw forecasts. This feature is a consequence of the reduction of the forcing biases, situation that will be further discussed in Sect. 3.5. The last feature is that the improvement of the forcings can help reducing the negative skill in streamflow forecasts. For example, April to November forecasts at longer lead times, generated with raw ECMWF System 4 forcings, exhibit a highly negative and statistically significant skill, sometimes lower than -1.0. Streamflow forecasts generated with preprocessed forcings for those months tend to have a neutral skill. This in turn ~~implying~~ implies that their accuracy is not different from the accuracy of ESP~~obtained with ESP~~ forecasts. The final conclusion is related to sharpness. As we can see in Fig. 5, streamflow forecasts generated with preprocessed forcings have an ensemble range that is wider than the reference forecasts.

The second and third columns in Fig. 6 show the PIT diagrams of streamflow forecasts generated with preprocessed forcings for the winter and summer forecasts at the first month lead time. The statistical consistency for the winter months seems to be ~~worsening~~worse than~~, in comparison to~~ the consistency of ~~the~~ forecasts generated with raw forcings. The same degree of deterioration is seen for both preprocessing methods. This is caused by compensational errors that will be further discussed in Sect. 3.5. Besides from that particular season, improvements in consistency after preprocessing can be seen during the autumn (not shown), and August, although to a lesser degree. For spring and early summer forecasts, the same level of consistency is observed for both the raw and preprocessed forecasts. At longer lead times, the benefit of preprocessing for statistical consistency is clearer, most of the months pass the uniformity test.

3.4 Postprocessed streamflow forecasts

The final step in the analysis is the postprocessing of streamflow forecasts generated with raw and preprocessed ECMWF System 4 forcings. Fig. 7 shows the verification results that can be directly compared to the results in Fig. 5.

The first column in Fig. 7 shows a clear reduction of the absolute bias compared to the raw and preprocessed generated forecasts. Bias lies within the range [-10,10%], for all months and lead times. Furthermore, the majority of the CRPSS for all months and lead times are positive, while a small negative skill is seen during the autumn. Note, however, that the differences in accuracy between ESP and the postprocessed forecasts are only significant at the 5% level for few target months and lead times. In general, there seems to be a worsening of the sharpness after postprocessing (Fig. 5). However, this deterioration is lower when comparing preprocessed versus postprocessed forecasts. Furthermore, the degree of the deterioration varies according to the target month. For example, summer months (June and July) exhibit a larger deterioration of sharpness, i.e., forecast spread is larger than that of the ESP. On the other hand, forecasts for late autumn and early December appear to be narrower than ESP forecasts after postprocessing.

Figure 8 shows the PIT diagrams for the summer and winter seasons at the first month lead time of the postprocessed streamflow forecasts. The plot can be directly compared to Fig. 6. As seen from the PIT diagram, all months in those seasons pass the uniformity test, indicating that after postprocessing, the observations can be considered as random samples of the predictive distribution. The remaining PIT diagrams for spring and autumn and lead times 2-7 months (not shown in Figure 8) show that statistical consistency is present for all months and lead times. At longer lead times, the CDFs of the z_i*'s* are closer to the 1:1 diagonal. This is achieved due to two factors: (i) the additional reduction of bias after postprocessing, and (ii) the worsening of sharpness for long lead times where the larger ensemble spread encloses a larger portion of observed values.

3.5 Effect of hydrological model bias in skill evaluations

As mentioned in Sect. 3.1., hydrological model biases, which are larger for the upstream station 21 (Fig. 4), combined with ~~the~~ structural biases in GCMs, can lead to a situation with a high skill resulting from compensational errors providing “the right forecast for the wrong reasons”. In order to illustrate this point, Fig. 9a-9b show the CRPSS for, respectively, station 21 with large bias (PBias = 48%, Fig. 4) and station 82 with small bias (PBias = 1.7%, Fig. 4). The figure shows CRPSS for forecasts generated with raw ECMWF forcings and preprocessed forcings with the LS method for the target months Jan-Dec at lead time 4 (e.g. January forecasts initiated in October). In addition to the comparison against observations, we also include a comparison against simulated streamflows (continuous run of the Ahlergaarde model with observed meteorological forcings, Fig. 4). This is done in order to remove the effect of hydrological model bias and hence focus the analyses on the biases coming from forcings alone.

The high skill against observed streamflows is more visible during the wettest months (November-April) for station 21 where hydrological model biases are highest (Fig. 4). Once the comparison is made against simulated streamflows, the high positive skill becomes highly negative (Fig. 9a). The deterioration of skill when compared against simulated streamflows is also seen at station 82 for Dec-March, although to a lesser extent (Fig. 9b). To illustrate why this happens, Fig. 9c and 9d show the monthly streamflow forecasts for all 24 years for target month December of forecasts initialized in September (lead time 4). Both ESP and raw (Fig. 9c) and preprocessed (Fig. 9d) forecasts are shown, along with their respective skill scores of accuracy, when the comparison is made against observed (CRPSS) and simulated (CRPSS.s) values.

Figure 9c shows two issues. First, the large hydrological model bias ~~that~~ cause ESP to have a deviation from the observations, leading to a high CRPS for the reference forecast in Eq. (4). Secondly, for the winter months, precipitation from the raw ECMWF System 4 forecasts exhibit a negative bias of around -~~15~~25% (Lucatero, et. al., ~~*in preparation*~~2017). This compensates the biased streamflow forecasts and results in a low $CRPS_{Sys}$ value in Eq. (4). The CRPSS then becomes positive and ~~high~~ large (0.54). However, when the comparison is done against simulated values, the skill score becomes highly negative (CRPSS.s = -0.41). Once the biases in the forcings are removed (Fig. 9d), then the hydrological model bias takes over, leading these forecasts to the same level as the ESP, increasing its CRPS, which in turn reduces the skill score (CRPSS = -0.04).

Note that the opposite situation arises, i. e. "the wrong forecast for the wrong reason", when hydrological model error is small and precipitation forecast bias is large. Biases in precipitation forecasts will propagate through streamflow forecasts, leading to streamflow bias of equal sign and of similar magnitude as precipitation bias. Streamflow bias is then reduced when meteorological forecast bias is removed (Fig. 5, second and third row). This situation appears during summer or autumn (Fig. 5, first row), when hydrological model error is smaller than in winter.

The apparent skill trend along target months of raw GCM-based streamflow forecasts (Fig. 5, first row) is a product the above explained error interactions rather than the existence (or lack) of predictability during the given months. Further analysis linking concurrent and/or previous hydrometeorological processes (i.e., accumulation of snowpack) to streamflow forecast skill would require additional research as discussed later. Moreover, preprocessed meteorological forecasts bias is invariant along lead time, and show only mild improvements in accuracy over ensemble climatology during the first month lead time (Lucatero, et. al., 2017). This situation, together with the reduction of error interactions negatively affecting streamflow forecasts at longer leads, produces the flattening of the trend in skill along lead time (Fig. 5, second and third row).

Stations like 21 could benefit the most from postprocessing ~~that will remove~~removing hydrological model biases that calibration alone could not remove. This is illustrated with the visualization of the CRPSS of the different forecasts in Fig. 10a-10d. The comparison is made against observations. Figure 10b shows a reduction of ~~the~~ skill after ~~the~~ raw forcings have been preprocessed, as a result of the compensation errors discussed above. However, once the hydrological biases are removed with postprocessing (Fig. 10c and 10d), skill is positive and significant throughout November to April. Note, however, that the high skill at this particular station is mainly driven by the poor performance of the reference ESP, due to the large bias of the hydrological model (Fig. 4). It is also worth noting the lack of differences in skill between Fig. 10c and Fig. 10d, showing that, for this particular location, a combination of preprocessing plus postprocessing is just as good as postprocessing of the forecasts generated with raw forcings alone.

3.6 Low flow forecasting

In addition to the evaluation of the monthly streamflow forecasts, we have assessed whether the use of GCM forecasts can add value to the forecasting of annual low flows compared to ~~the~~ ESP. Figure 11 shows ~~the~~ low flow forecast, issued in

Formatted: Font: Not Italic

April, at the outlet stations 82 for both the raw forecasts and the different pre- and postprocessing strategies. Black boxplots represent the forecast generated using the raw outputs of the ECMWF System 4 (Fig. 11a), the preprocessed forecasts (Fig. 11c and 11e) and the postprocessed forecasts (Fig. 11b, 11d, 11f). The box-plots in the background (blue) represent the ESP forecasts and the red dots represent the yearly observed minimum discharges. When we look at Fig. 10a, several features can be highlighted. First, despite the underprediction of the raw generated forecasts and, to a lesser extent, the ESP forecasts of the highest minimum discharges in the 00s, the year-to-year variability is replicated well, i.e., low observed low flow values have low, although biased, forecasted values, and high observed low flow values have, in general, high forecasted values. Secondly, even though the raw generated forecasts are sharper than the ESP by about 10% (SS = 0.11), they do not manage to beat ESP in terms of skill of accuracy (CRPSS = -0.14), i.e. they are overconfident.

Preprocessing meteorological forecasts seems to have a positive effect ~~in~~ on low flow forecasting, reducing the CRPSS from -0.14 to -0.01 when using the LS preprocessor. This happens because of the loss of sharpness (from 0.11 to -0.11) that allow the forecasts to better capture the higher low flows during the 00's. However, it is still difficult to beat the ESP. Postprocessing seems to have a similar effect, loss of sharpness, decrease in bias that allow the forecasts to capture the high low flows in the 00s and 10s. This situation, however, leads to a loss in skill in forecasting low flows in the 90s, leveling out the skill to a similar score (CRPSS = -0.12) as the forecasts generated with raw ECMWF forcings (CRPSS = -0.14). ~~It thus~~ Thus, it seems that an attempt to reduce meteorological and hydrological biases through processing the forcings and/or the streamflow will result in only a modest increase in skill of low flow predictions on average. ESP remains a reference forecast system difficult to beat.

4 ~~Summary and conclusions~~ Discussion

~~Seasonal forecasts of streamflows initiated in each calendar month for the 1990-2013 period were generated for a groundwater dominated catchment located in a region where seasonal atmospheric forecasting is a challenge. We analyzed the bias and statistical consistency of monthly streamflow forecasts forced with ECMWF System 4 seasonal forecasts along all calendar months throughout the year. In addition to this, we evaluated their accuracy and sharpness relative to that of the forecasts generated with an ensemble of historical meteorological observations, the ESP. Due to the systematic errors of GCM seasonal forecasts and errors in the hydrological model that calibration alone cannot defuse, both pre and postprocessing using two popular and simple correction techniques were used to remove them: LS and QM. Finally, we also estimate the skill that the different forecast generation approaches have on forecasting the minimum yearly discharge.~~

Monthly streamflow forecasts derived for raw, preprocessed meteorology and postprocessed streamflow, in general show limited skill beyond one month lead time. This is not a surprising result, given the limited capabilities of skill of meteorological forecasts in the region Lucatero, et al. (2017). This result has been documented in Wood et al., (2005); Yuan et al., (2013) and more recently Crochemore, et al. (2016) for a region near to our focus region study. GCM-based streamflow forecasts could be then of potential use if the end user is interested in gaining accuracy of forecasts for the next month only. Moreover, we were able to demonstrate that, at least for a groundwater dominated catchment located in a region with temperate climate, GCM ability to improve low flow forecasts is also limited, regardless of any attempts to correct forcings and/or streamflow forecasts. Further research could focus on the usefulness of GCM forecasts for drought forecasting, i.e., magnitude, duration and severity (Fundel, 2013) in comparison to forecasts generated with the ESP method.

Furthermore, caution must be taken when hydrological model errors are large, as it may lead to erroneous evaluations of skill when hydrological model biases are neutralized by opposite GCM errors, e.g., forecasts of monthly streamflow during the winter. This is an issue somewhat underexplored in studies of forecast skill and should be evaluated especially when calibration objective function differs from the final forecast quantity of interest, and when no attempts to remove biases in meteorological forecasts is made.

Preprocessing of the forcings alone helped to reduce streamflow biases and reduce the negative skill at longer lead times. The reduction of the under- or overestimation, lead to forecasts with a higher statistical consistency, for most of the months and lead times considered. This rather mild enhancement was also found by Crochemore et al., (2016). Moreover, postprocessing alone does a better job in removing biases in the mean, which in turn helps to ameliorate issues with the statistical consistency. Ye et al., (2015) and Zalachori et al., (2012) also report the above behaviour, whereas Yuan and Wood, (2012) found a better correction of statistical consistency after both pre- and postprocessing. The removal of biases of both forcings and hydrological model did not ensure a higher level of accuracy than the ESP, as demonstrated by the non-significant differences of accuracy between GCM-based forecasts and the ESP. This is also true for the low flow forecasts as mentioned above.

The methods used here for preprocessing (LS and QM) and posprocessing (QM) were chosen because of their simplicity. However, postprocessing in general is a field that has been gaining traction over the last decade with a variety of methods that differ in their mathematical sophistication. The reader is referred to Li (2017) for a detailed and updated literature review on the subject. Moreover, QM disadvantages have been widely discussed in Zhao, (2017) and references therein. The main issue concerning the fact that when forecast-observation linear relationship is weak or non-existent, QM has difficulties creating forecasts that are consistent (i.e. skill at least as good as the reference forecast). Other methods could have been used that allow for correction of both statistical consistency together with consistency. However, the benefits of the more sophisticated methods might be dampened due to limited sample size, which is often the case in hydrometeorological forecasting. Nevertheless, our present study could be extended by analyzing the added skill gained by the increased complexity of processing methods, using the same reforecast dataset, such as the case of Mendoza, (2017) although its application focused on statistical forecasting.

Formatted: Font:

Another obvious omission of the study is the exploitation of storages in the form of snow, soil moisture or/and groundwater and taking advantage of the hydrological memory that may increase skill at longer leads. This has been the routine for snow dominated catchments in western U. S. by means of ESP (Wood and Lettenmaier, 2006). However, a preliminary evaluation of the relationship between groundwater levels in winter and low flows during the summer in the Ahlergaarde catchment studied here showed that relatively high correlations exist in large parts of the catchment (Kidmose, et al., *unpublished work*). This correlation can be further explored in forecasting mode to extend positive skill lead time by means of data assimilation (Zhang et al., 2016) or by statistical postprocessing of streamflow forecasts (Mendoza, et al., 2017). Moreover, predictability attribution studies exist that quantify the sensitivity of the skill of a forecasting system relative to different degrees of uncertainty, either in the forcing or the initial conditions. Wood et al., (2016) developed a framework to detect where to concentrate on improvements, e.g., either the initial conditions, usually by means of data assimilation (Zhang et al., 2016), or the seasonal climate forecasts. This might shed light on, and possibly reinforce the hypothesis that for groundwater dominated catchments and forecasting of low flows, initial conditions will have a higher influence on forecast skill at longer lead times (Paiva et al., 2012; Fundel et al., 2013).

~~Monthly streamflow forecasts generated with raw forcings are affected by biases both in the hydrological model and the forcing. This bias grows as lead time increases and also reflected in the shape of the PIT diagrams that exhibit a persistent underestimation of the forecast mean for the majority of the calendar months and lead times. GCM based forecasts are, however, sharper than ESP forecasts, which combined with the high biases lead to a poorer accuracy than ESP forecasts. Our results are directly comparable to those of Crochemore et al., (2016) as they also deal with seasonal streamflow forecasting with ECMWF System 4 for a region located not too far from our study area. They also found GCM based forecast to be sharper than those of the ESP. However, biases intrinsic in the GCM are also propagated into the streamflow forecasts,~~

~~which lead them to have a poorer accuracy. Overall, the GCM-based forecasts did not performed better that those based on ESP. Other studies that have compared the skill of GCM-based streamflow forecasts versus those based on ESP have found similar results (Wood et al., 2005; Yuan et al., 2013). In the present study we found that this is especially difficult during the dry months and for streamflow stations that exhibit low hydrological model error. Furthermore, caution must be taken when hydrological model errors are high, as it may lead to erroneous evaluations of skill when hydrological model biases are neutralized by opposite GCM errors, e.g., forecasts of monthly streamflow during the winter.~~

~~Preprocessing of the forcings alone helped to reduce streamflow biases and reduce the negative skill at longer lead times. The reduction of the under- or overestimation, lead to forecasts with a higher statistical consistency, for most of the months and lead times considered. This rather mild enhancement was also found by Crochemore et al., (2016). Moreover, postprocessing alone does a better job in removing biases in the mean, that in turn help to ameliorate issues with the statistical consistency. Ye et al., (2015) and Zalachori et al., (2012) also report the above behaviour, whereas Yuan and Wood, (2012) found a better correction of statistical consistency after both pre- and postprocessing. The removal of biases of both forcings and hydrological model did not ensure a higher level of accuracy than the ESP, as demonstrated by the non-significant differences of accuracy between GCM-based forecasts and the ESP. This is also true for the low flow forecasts as mentioned above.~~

~~One must keep in mind, however, that the results presented here may depend on the catchment characteristics as well as on the climatic conditions of the study area. Seasonal meteorological forecasting is still a difficult task, particularly on regions further away from the tropics, which in turn translate into the streamflow forecasts. One obvious omission of the study presented is the exploitation of storages in the form of snow, soil moisture or/and groundwater and taking advantage of the hydrological memory that may increase skill at longer leads. This has been the routine for snow dominated catchments in western U. S. by means of ESP (Wood and Lettenmaier, 2006). However, a preliminary evaluation of the relationship between groundwater levels in winter and low flows during the summer in the same catchment studied here showed that relatively high correlations exist in large parts of the catchment (Kidmose, et al., unpublished work). Predictability attribution studies exist that quantify the sensitivity of the skill of a forecasting system relative to different degrees of uncertainty, either in the forcing or the initial conditions. Wood et al., (2016) developed a framework to detect where to concentrate on improvements, e.g., either the initial conditions, usually by means of data assimilation (Zhang et al., 2016), or the seasonal climate forecasts. This might shed light on, and possibly reinforce the hypothesis that for groundwater dominated catchments and forecasting of low flows, initial conditions will have a higher influence on forecast skill at longer lead times (Paiva et al., 2012).~~

5 Conclusion

Seasonal forecasts of streamflows initiated in each calendar month for the 1990-2013 period were generated for a groundwater dominated catchment located in a region where seasonal atmospheric forecasting is a challenge. We analyzed the bias and statistical consistency of monthly streamflow forecasts forced with ECMWF System 4 seasonal forecasts along all calendar months throughout the year. In addition to this, we evaluated their accuracy and sharpness relative to that of the forecasts generated with an ensemble of historical meteorological observations, the ESP. Monthly streamflow forecasts generated with raw ECMWF System 4 forcings show skill only during the winter months at the first month lead time. Nevertheless, it was shown that the apparent large skill is an effect of compensational errors between meteorological forecasts and hydrological model. Due to ~~the systematic errors~~biases of GCM-based meteorological seasonal forecasts and errors in the hydrological model that calibration alone cannot defuse, both pre and postprocessing using two popular and simple correction techniques were used to remove them: LS and QM. Finally, we also estimate the skill that the different forecast generation approaches have on forecasting the minimum yearly discharge. Our results show that postprocessing streamflow allows for the most gain in skill and statistical consistency. However, monthly streamflow and low flow forecasts generated with forcings from GCM still show difficulties beating ESP forecasts, especially at lead times longer than 1 month.

~~5 Appendix A. Interpretation of the PIT Diagram~~

~~Fig. A1 serves as a basis for the interpretation of the PIT diagrams. The figure is modified from Laio and Tamea, (2007). Four situations can arise: (i) overprediction, or positive bias in the mean when the CDF of the z_i's lies above the 1:1 diagonal; (ii) underprediction, or negative bias in the mean when the CDF of the z_i's lies below the 1:1 diagonal; (iii) overdispersion, or underconfident forecasts (large forecasts) when a greater proportion of the values of the CDF lie on the middle ranges bins of the distribution; and (iv) underdispersion, or negative bias in spread (overconfidence) when 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 Andy Wood and Pablo Mendoza for hosting the first author at NCAR. We also express our gratitude to Massimiliano Zappa and one anonymous reviewer for their comments that improved the quality of this paper.

Field Code Changed

6 References

Bogner, K., Liechti, K. and Zappa, M.: Post-processing of stream flows in Switzerland with an emphasis on low flows and floods, Water (Switzerland), 8(4), doi:10.3390/w8040115, 2016.

Bruno Soares, M. and Dessai, S.: Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe, Clim. Change, 137(1–2), 89–103, doi:10.1007/s10584-016-1671-8, 2016.

Córdoba-Machado, S., Palomino-Lemus, R., Gámiz-Fortis, S. R., Castro-Díez, Y. and Esteban-Parra, M. J.: Seasonal streamflow prediction in Colombia using atmospheric and oceanic patterns, J. Hydrol., 538, doi:10.1016/j.jhydrol.2016.04.003, 2016.

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.

D'Agostino, R. B. and Stephens, A. M., eds.: Goodness-of-fit techniques, Dekker, New York, 1986.

Day GN. 1985. Extended stream flow forecasting Using NWSRFS, J. Water Resour. Plann. Manage., 111(2): 157–170.

Demirel, M. C., Booij, M. J. and Hoekstra, a. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, Hydrol. Earth Syst. Sci., 19(1), 275–291, doi:10.5194/hess-19-275-2015, 2015.

Doherty, J.: PEST Model-Independent Parameter Estimation User Manual Part I: PEST, SENSAN and Global Optimisers, Watermark Numer. Comput., 279, 2016.

Friederichs, P. and Thorarinsdottir, T. L.: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, Environmetrics, 23(7), 579–594, doi:10.1002/env.2176, 2012.

Fundel, F., Jörg-Hess, S. and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, Hydrol. Earth Syst. Sci., 17(1), 395–407, doi:10.5194/hess-17-395-2013, 2013.

Graham, D. N. and Butts, M. B.: Flexible, integrated watershed modelling with MIKE SHE. In Watershed Models, Eds. V.P. Singh & D.K. Frevert Pages 245-272, CRC Press. ISBN: 0849336090., 2005.

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.

Jaun, S., Ahrens, B., Walser, A., Ewen, T. and Schär, C.: A probabilistic view on the August 2005 floods in the upper Rhine catchment, Nat. Hazards Earth Syst. Sci., 8(2), 281–291, doi:10.5194/nhess-8-281-2008, 2008.

Jensen, K. H. and Illangasekare, T. H.: HOBE: A Hydrological Observatory, Vadose Zo. J., 10, 1–7, doi:10.2136/vzj2011.0006, 2011.

Kidmose, J., Refsgaard, J.C., Troldborg, L. and Stisen, S.: Seasonal memory of a hydrological groundwater - surface water system, *in preparation*.

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

Lucatero, D., Madsen , H., Refsgaard, J.C., Kidmose, J., and Jensen, K. H.: On the skill of raw and postprocessed ensemble seasonal meteorological forecasts in Denmark, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10. 5194/hess-2017-366, 2017.

Madadgar, S., Moradkhani, H. and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrol. Process., 28(1), 104–122, doi:10.1002/hyp.9562, 2014.

Mendoza, P. A., Wood, A. W., Clark, E., Nijssen, B., Clark, M. P., Rothwell, E., Arnold, J. R. and Brekke, L. D.: An intercomparison of approaches for harnessing sources of predictability in operational seasonal streamflow forecasting, Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-60, 2017.

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 Technical Memorandum 656, November, 49, 2011.

Olsson, J., Uvo, C. B., Foster, K. and Yang, W.: Technical Note: Initial assessment of a multi-method approach to spring-flood forecasting in Sweden, Hydrol. Earth Syst. Sci., 20(2), 659–667, doi:10.5194/hess-20-659-2016, 2016.

Paiva, R. C. D., Collischonn, W., Bonnet, M. P. and De Gonçalves, L. G. G.: On the sources of hydrological prediction

Formatted: English (United States)

uncertainty in the Amazon, Hydrol. Earth Syst. Sci., 16(9), 3127–3137, doi:10.5194/hess-16-3127-2012, 2012.

Robertson, D. E., Pokhrel, P. and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrol. Earth Syst. Sci., 17(2), 579–593, doi:10.5194/hess-17-579-2013, 2013.

Robertson, D. E. and Wang, Q. J.: A Bayesian Approach to Predictor Selection for Seasonal Streamflow Forecasting, J. Hydrometeorol., 13(1), 155–171, doi:10.1175/JHM-D-10-05009.1, 2012.

Rosenberg, E. A., Wood, A. W. and Steinemann, A. C.: Statistical applications of physically based hydrologic models to seasonal streamflow forecasts, Water Resour. Res., 47(9), doi:10.1029/2010WR010101, 2011.

Roulin, E. and Vannitsem, S.: Post-processing of medium-range probabilistic hydrological forecasting: Impact of forcing, initial conditions and model errors, Hydrol. Process., 29(6), 1434–1449, doi:10.1002/hyp.10259, 2015.

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.

Schepen, A., Zhao, T., Wang, Q. J., Zhou, S. and Feikema, P.: Optimising seasonal streamflow forecast lead time for operational decision making in Australia, Hydrol. Earth Syst. Sci., 20(10), 4117–4128, doi:10.5194/hess-20-4117-2016, 2016.

Shamir, E.: The value and skill of seasonal forecasts for water resources management in the Upper Santa Cruz River basin, southern Arizona, J. Arid Environ., 137, 35–45, doi:10.1016/j.jaridenv.2016.10.011, 2017.

Shi, X., Wood, A. W. and Lettenmaier, D. P.: How Essential is Hydrologic Model Calibration to Seasonal Streamflow Forecasting?, J. Hydrometeorol., 9, 1350–1363, doi:10.1175/2008JHM1001.1, 2008.

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.

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

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.

Verkade, J. S., Brown, J. D., Reggiani, P. and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, doi:10.1016/j.jhydrol.2013.07.039, 2013.

Wang, Q. J., Robertson, D. E. and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resour. Res., 45(5), 1–18, doi:10.1029/2008WR007355, 2009.

Wood, A. W., Maurer, E. P., Kumar, A. and Lettenmaier, D. P.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res. D Atmos., 107(20), 1–15, doi:10.1029/2001JD000659, 2002.

Wood, A. W., Kumar, A. and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental prediction climate model-based ensemble hydrologic forecasting in the western United States, J. Geophys. Res. D Atmos., 110(4), 1–16, doi:10.1029/2004JD004508, 2005.

Wood, A. W. and Lettenmaier, D. P.: A test bed for new seasonal hydrologic forecasting approaches in the western United States, Bull. Am. Meteorol. Soc., 87(12), 1699–1712, doi:10.1175/BAMS-87-12-1699, 2006.

Wood, A. W. and Schaake, J. C.: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, J. Hydrometeorol., 9(1), 132–148, doi:10.1175/2007JHM862.1, 2008.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J. and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, J. Hydrometeorol., 17(2), 651–668, doi:10.1175/JHM-D-14-0213.1, 2016.

Yapo, P. O., Gupta, H. V. and Sorooshian, S.: Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, J. Hydrol., 181(1–4), 23–48, doi:10.1016/0022-1694(95)02918-4, 1996.

Ye, A., Duan, Q., Schaake, J., Xu, J., Deng, X., Di, Z., Miao, C. and Gong, W.: Post-processing of ensemble forecasts in low-flow period, Hydrol. Process., 29(10), 2438–2453, doi:10.1002/hyp.10374, 2015.

Yuan, X., Wood, E. F., Luo, L. and Pan, M.: A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction, Geophys. Res. Lett., 38(13), 1–7, doi:10.1029/2011GL047792, 2011.

Yuan, X. and Wood, E. F.: Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast, Water Resour. Res., 48(12), 1–7, doi:10.1029/2012WR012256, 2012.

Yuan, X., Wood, E. F., Roundy, J. K. and Pan, M.: CFSv2-Based seasonal hydroclimatic forecasts over the conterminous United States, J. Clim., 26(13), 4828–4847, doi:10.1175/JCLI-D-12-00683.1, 2013.

Yuan, X., Roundy, J. K., Wood, E. F. and Sheffield, J.: Seasonal forecasting of global hydrologic extremes : System development and evaluation over GEWEX basins, Bull. Am. Meteorol. Soc., 96(11), 1895–1912, doi:10.1175/BAMS-D-14-00003.1, 2015.

Yuan, X.: An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 2: The added value from climate forecast models, Hydrol. Earth Syst. Sci., 20(6), 2453–2466, doi:10.5194/hess-20-2453-2016, 2016.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T. and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Adv. Sci. Res., 8, 135–141, doi:10.5194/asr-8-135-2012, 2012.

Zhang, D., Madsen, H., Ridler, M. E., Kidmose, J., Jensen, K. H. and Refsgaard, J. C.: Multivariate hydrological data assimilation of soil moisture and groundwater head, Hydrol. Earth Syst. Sci., 20(10), 4341–4357, doi:10.5194/hess-20-4341-2016, 2016.

Zhao, L., Duan, Q., Schaake, J., Ye, A. and Xia, J.: A hydrologic post-processor for ensemble streamflow predictions, Adv. Geosci., 29(1999), 51–59, doi:10.5194/adgeo-29-51-2011, 2011.

~~8~~ 7 Figure Captions

Figure 1: Diagram of generation of forecasts and verification procedures. RAW refers to the uncorrected ECMWF System 4 forecasts, while LS and QM refer to forecasts (either meteorological or hydrological) that are corrected using the Linear Scaling/Delta Change or Quantile Mapping method, respectively, for precipitation (P), temperature (T) and reference evapotranspiration (E_{T0}).

Figure 2: Location and topography of the Ahlergaarde catchment. The outlet station (82) and the upstream sub-catchment (21) are used in the study.

Figure 3: Climatology of the Ahlergaarde catchment. Values for precipitation (P), reference evapotranspiration (E_{T0}), streamflow (Q) and temperature (T) are monthly average values over the period 1990-2013.

Figure 4: Hydrographs for the 2000-2003 period. Percentage bias (PBias) and Nash-Sutcliffe efficiency score (NSE) are computed using the daily observed-simulated values for the complete 1990-2013 period. The scatter plots represent the observed-simulated annual low flow values.

Figure 5: PBias and skill in terms of accuracy and sharpness of monthly means of daily streamflow of raw and preprocessed forecasts at station 82. The 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 show a positive bias/skill and values in red a negative bias/skill. LS and QM represent the bias for a streamflow forecast generated with bias adjusted forcing using the Linear Scaling-Delta Change and Quantile Mapping, respectively. Circles represent the cases where the distribution of the accuracy and/or sharpness for ESP differs from that of the ECMWF System 4 generated forecasts at a 5% significance level using the WMW-test.

Figure 6: Statistical consistency of monthly mean of daily streamflow forecasts for winter (upper row) and summer (bottom row) for station 82. LS and QM represent the consistency of a streamflow forecast generated with bias adjusted forcing using the Linear Scaling and Quantile Mapping methods, respectively. Different colors represent different months in the season. The black lines parallel to the 1.1 diagonal are the Kolmogorov bands at the 5% significance level.

Figure 7: PBias and skill (sharpness and accuracy) of daily monthly mean streamflow forecasts for a post-processed forecast using the QM method for predictions generated using raw and preprocessed forcings from ECMWF System 4. Legend is the same as Figure 5.

Figure 8: Statistical consistency of daily monthly mean streamflow postprocessed forecasts for summer (upper row) and winter (bottom row) for station 82. The black lines parallel to the 1.1 diagonal are the Kolmogorov bands at the 5% significance level.

Figure 9: (a) and (b) Skill of accuracy (CRPSS) for upstream station 21 and outlet station 82 for target months Jan-Dec at lead time 4. Triangles and circles represent the forecasts generated with raw ECMWF System 4 forcings and preprocessed with LS, respectively, whereas black and blue lines represent the comparison against observed and simulated streamflow, respectively. The second row shows the monthly forecasts of December streamflow initialized in September (4 month lead time) for predictions using raw (c) and preprocessed (d) forcings for all years in the record (1990-2013) for station 21.

Figure 10: CRPSS of station 21 for forecasts generated with raw (a) and preprocessed (b) forcings, in addition to the postprocessed forecasts (c-d).

Figure 11: Low flow forecasts for the years 1990-2013 for generated using raw forcings (a), preprocessed forcings with the LS (c) and the QM (e) method and postprocessed streamflow for forecasts generated with raw (b) and preprocessed inputs (d

and f). The forecasts are initiated in April. Blue shaded box-plots are ESP forecasts. CRPSS and SS are computed using Eq. (4) with ESP as reference.

~~**Figure A1:** Interpretation of the Probabilistic Integral Transform Diagram (PIT). Observed values are generated with a standard normal distribution~~ $N(0,1)$~~. Instances of bias in both mean and dispersion are generated with the following distributions~~ $N(1.5,1)$~~,~~ $N(-1.5,1)$~~,~~ $N(0,3)$ ~~and~~ $N(0,0.3)$ ~~for an overestimated, underestimated, overdispersive and underdispersive system, respectively.~~

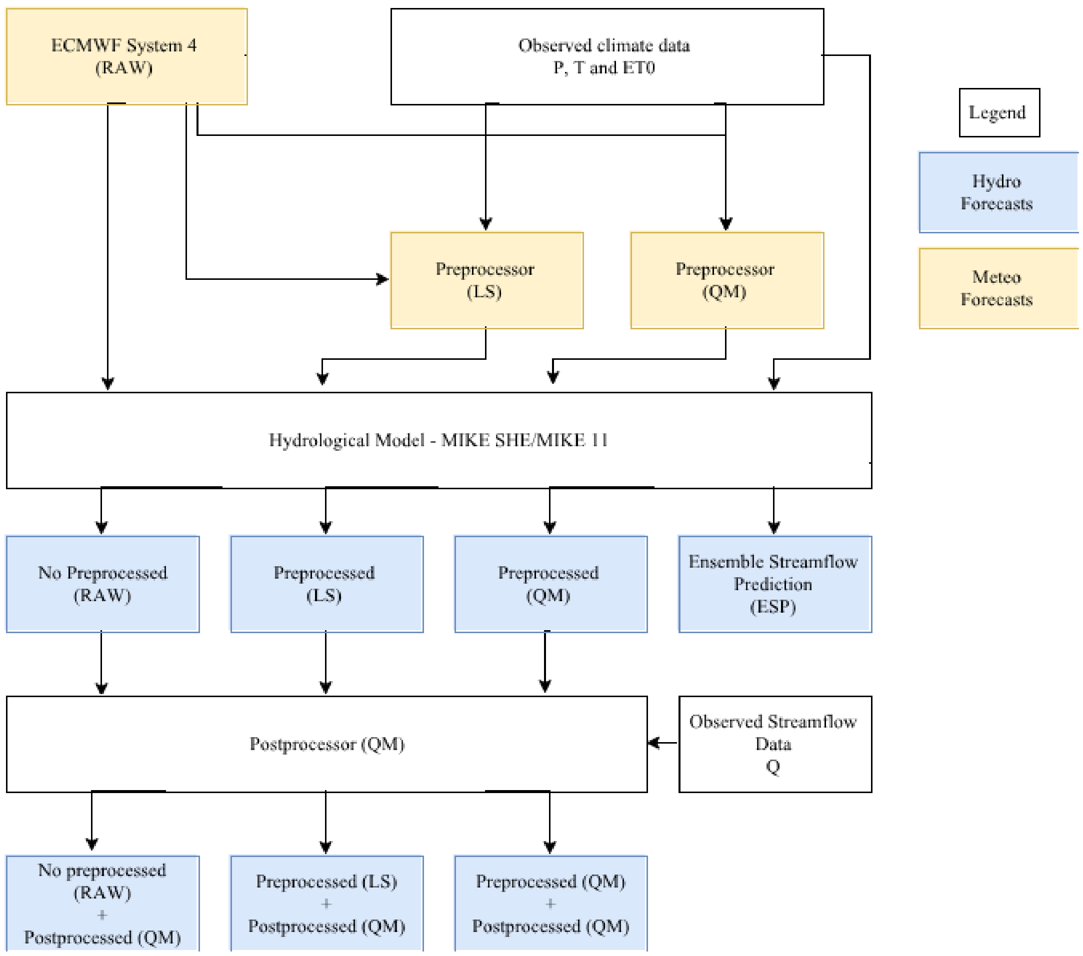


Figure 1: Diagram of generation of forecasts and verification procedures. RAW refers to the uncorrected ECMWF System 4 forecasts, while LS and QM refer to forecasts (either meteorological or hydrological) that are corrected using the Linear Scaling/Delta Change or Quantile Mapping method, respectively, for precipitation (P), temperature (T) and reference evapotranspiration (E_{T0}). Preprocessed refers to streamflow forecast generated with corrected meteorological forecasts while postprocessed refers to corrected streamflow forecasts.

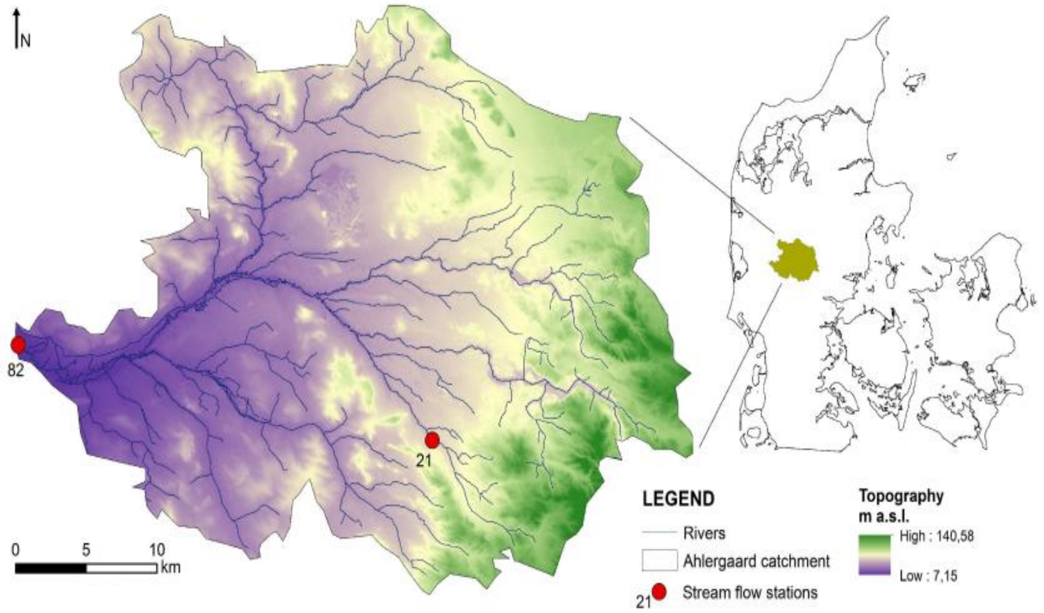


Figure 2: Location and topography of the Ahlergaarde catchment. The outlet station (82) and the upstream sub-catchment (21) are used in the study.

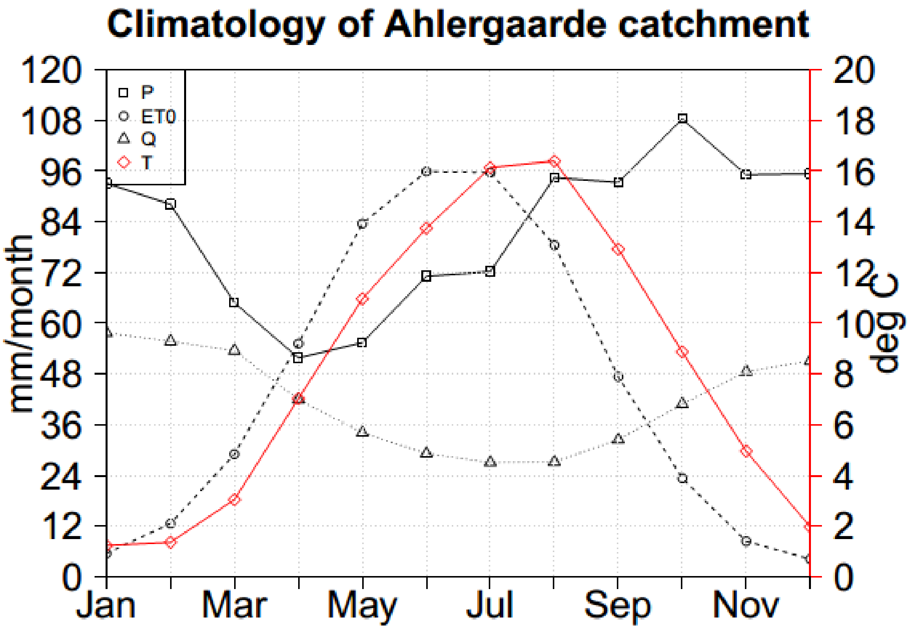


Figure 3: Climatology of the Ahlegaarde catchment. Values for precipitation (P), reference evapotranspiration (E_{T0}), streamflow (Q) and temperature (T) are monthly average values over the period 1990-2013.

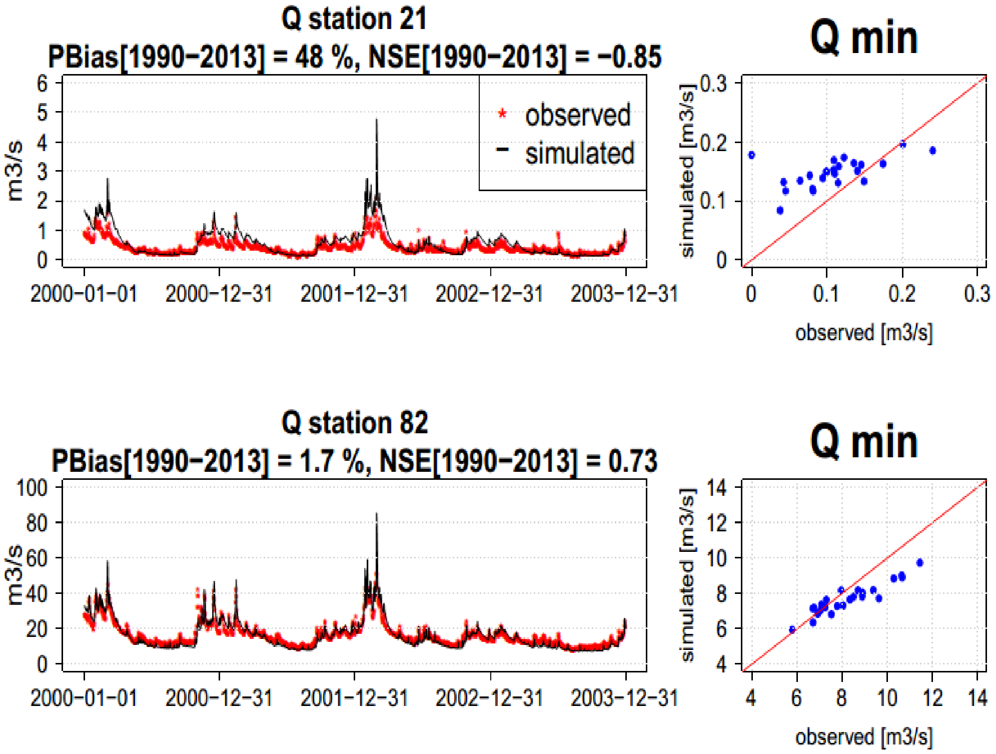


Figure 4: Hydrographs for the 2000-2003 period. Percentage bias (PBias) and Nash-Sutcliffe efficiency score (NSE) are computed using the daily observed-simulated values for the complete 1990-2013 period. The scatter plots represent the observed-simulated annual low flow values.

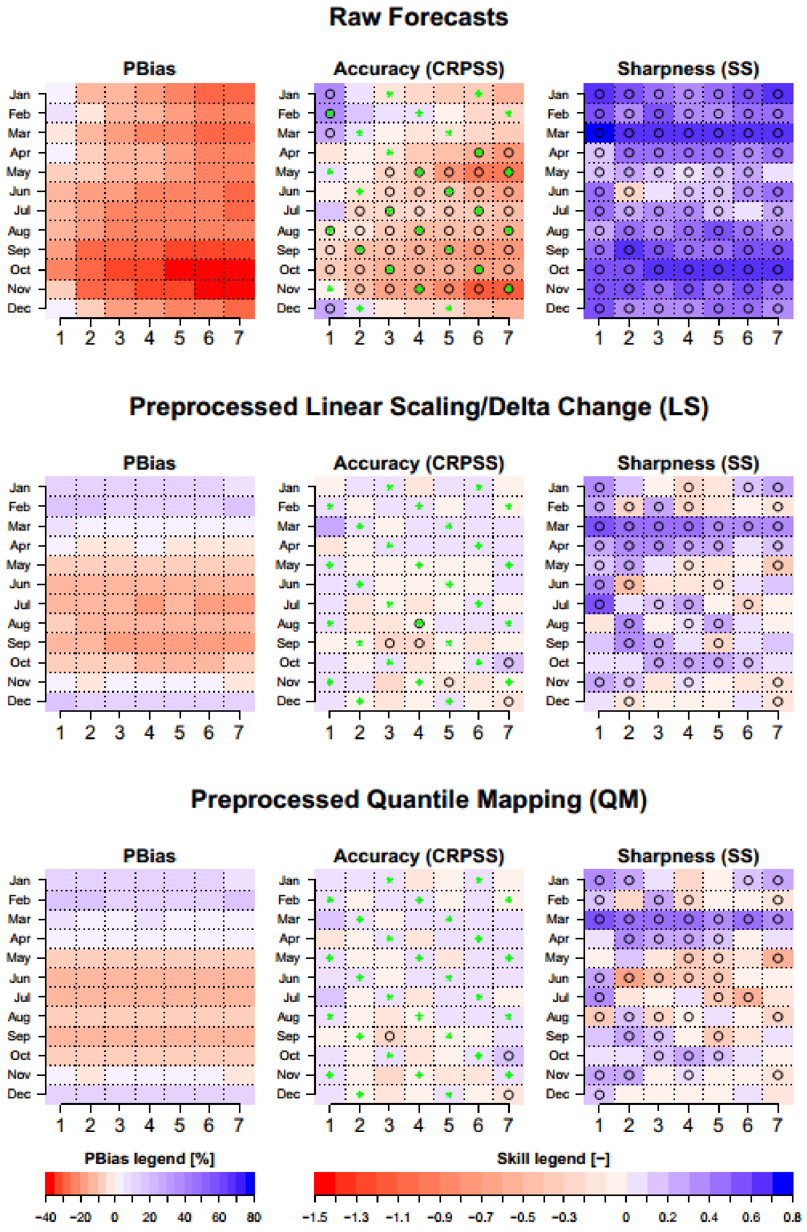


Figure 5: PBias and skill in terms of accuracy and sharpness of monthly means of daily streamflow of raw and preprocessed forecasts at station 82. The Y-axis represents the target month, and the X-axis represents the different lead times at which target months are forecasted. Values in blue range show a positive bias/skill and values in red a negative bias/skill. LS and QM represent the bias for a streamflow forecast generated with bias adjusted forcing using the Linear Scaling-Delta Change and Quantile Mapping, respectively. Circles represent the cases where the distribution of the accuracy and/or sharpness for ESP differs from that of the ECMWF System 4 generated forecasts at a 5% significance level using the WMW-test. Green crosses represent the months/lead times were the ensemble size is 51.

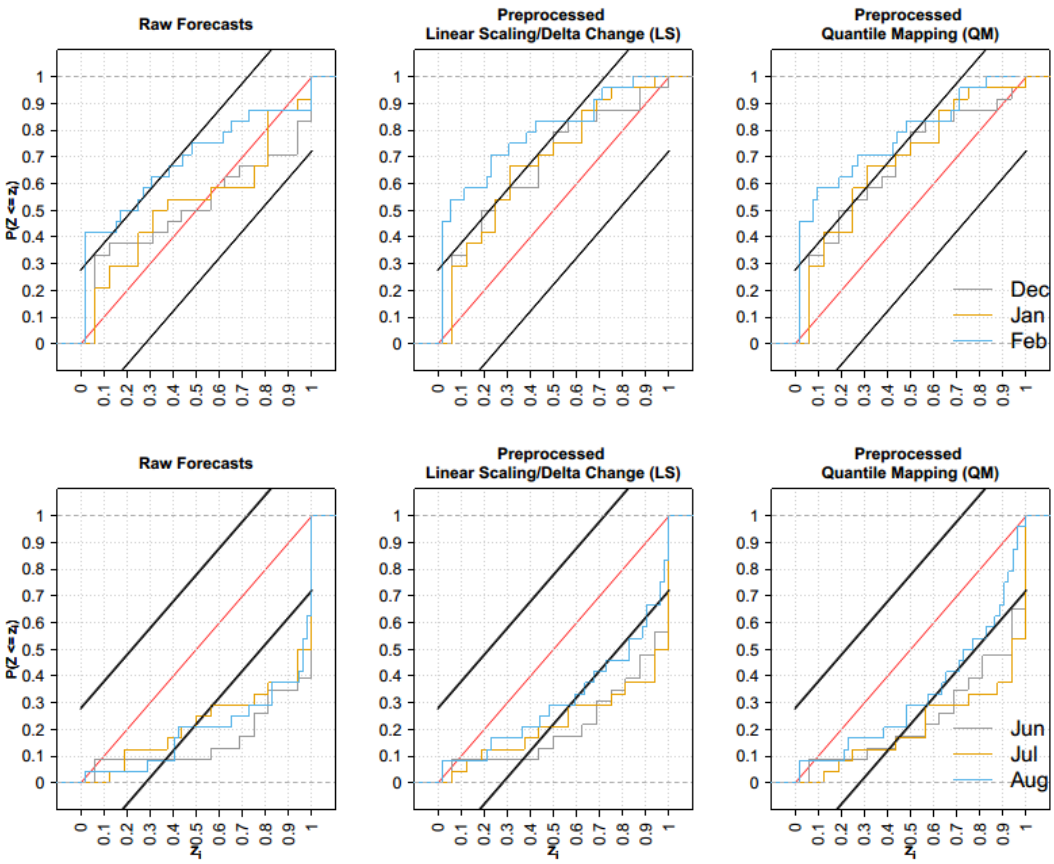


Figure 6: ~~Statistical consistency~~PIT diagrams of monthly mean of daily streamflow forecasts for winter (upper row) and summer (bottom row) for station 82. LS and QM represent the consistency of a streamflow forecast generated with bias adjusted forcing using the Linear Scaling and Quantile Mapping methods, respectively. Different colors represent different months in the season. The black lines parallel to the 1.1 diagonal are the Kolmogorov bands at the 5% significance level.

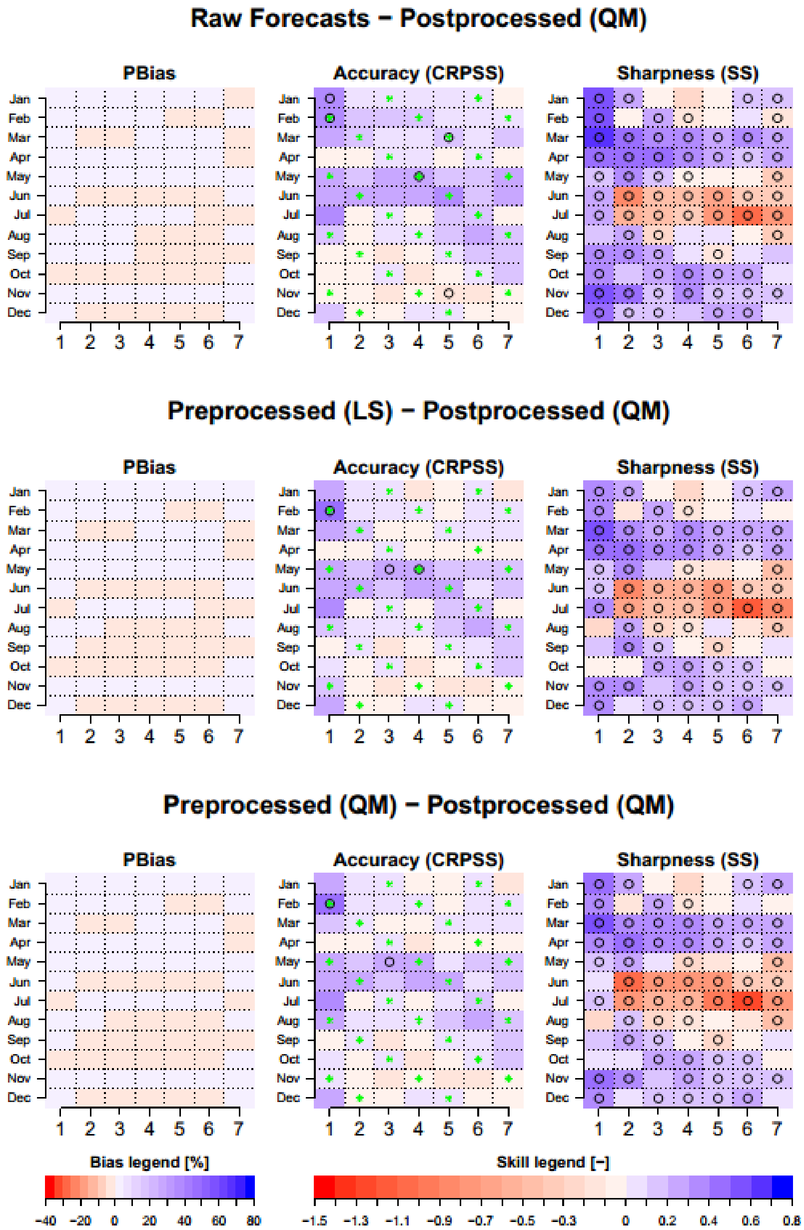


Figure 7: PBias and skill (sharpness and accuracy) of daily monthly mean streamflow forecasts for a post-processed forecast using the QM method for predictions generated using raw and preprocessed forcings from ECMWF System 4. Legend is the same as Figure 5. Green crosses represent the months/lead times were the ensemble size is 51.

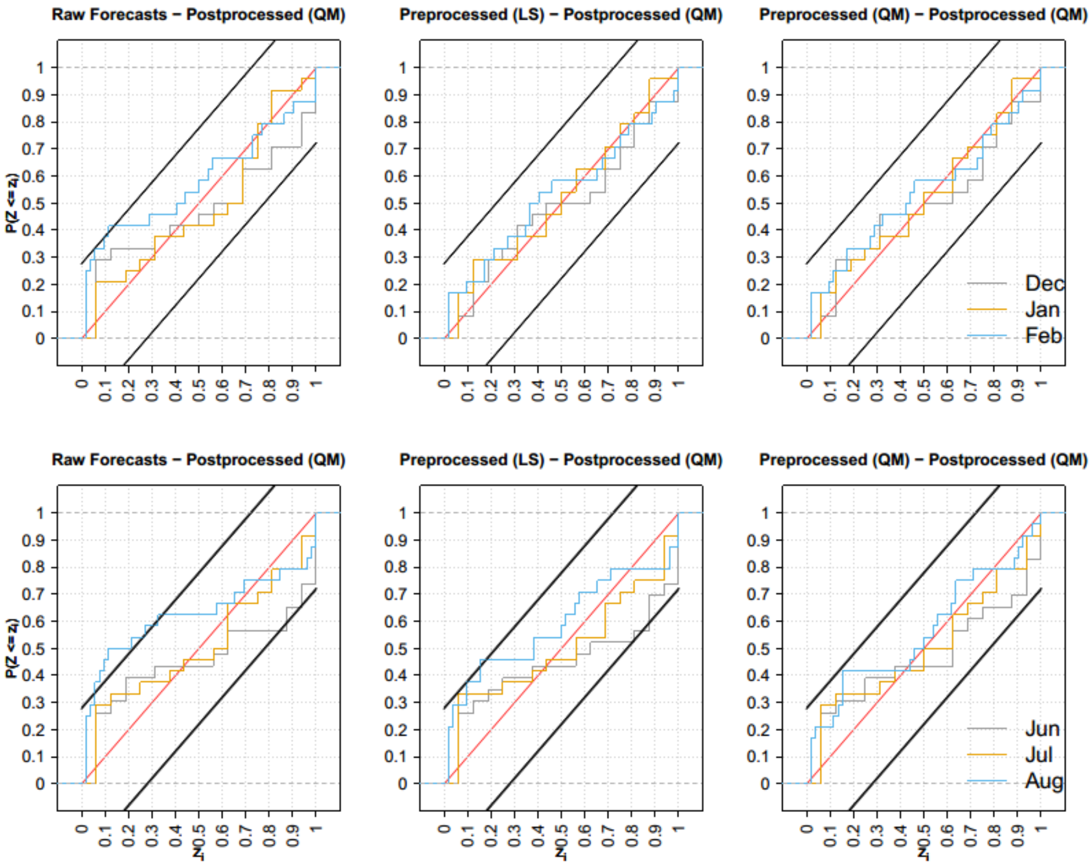


Figure 8: ~~Statistical consistency~~PIT diagrams of daily monthly mean streamflow postprocessed forecasts for summer (upper row) and winter (bottom row) for station 82. The black lines parallel to the 1.1 diagonal are the Kolmogorov bands at the 5% significance level.

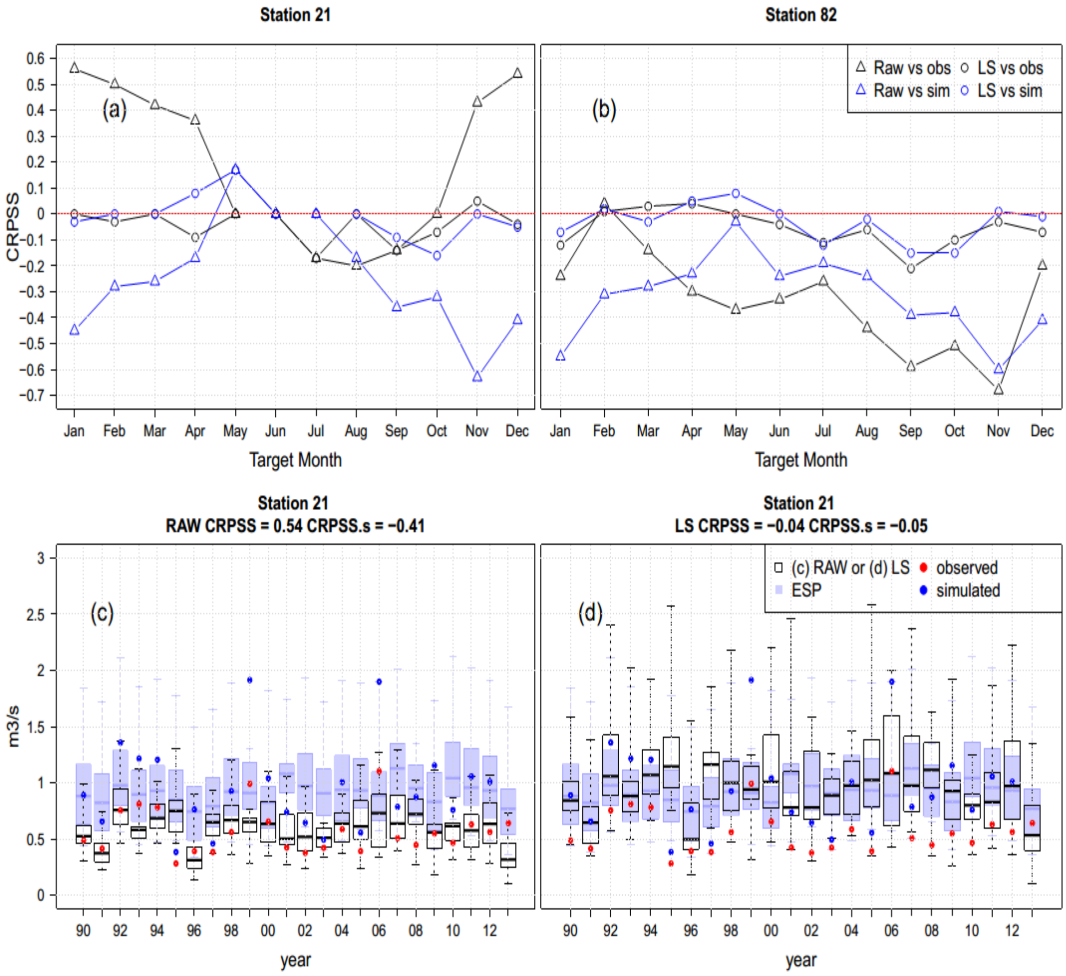


Figure 9: (a) and (b) Skill of accuracy (CRPSS) for upstream station 21 and outlet station 82 for target months Jan-Dec at lead time 4. Triangles and circles represent the forecasts generated with raw ECMWF System 4 forcings and preprocessed with LS, respectively, whereas black and blue lines represent the comparison against observed and simulated streamflow, respectively. The second row shows the monthly forecasts of December streamflow initialized in September (4 month lead time) for predictions using raw (c) and preprocessed (d) forcings for all years in the record (1990-2013) for station 21.

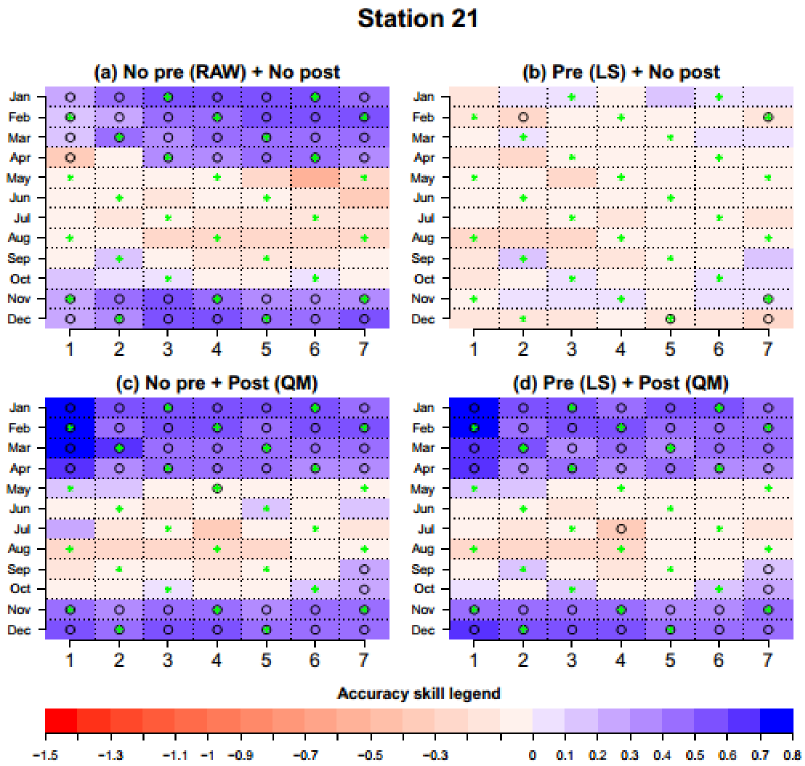


Figure 10: CRPSS of station 21 for forecasts generated with raw (a) and preprocessed (b) forcings, in addition to the postprocessed forecasts (c-d).

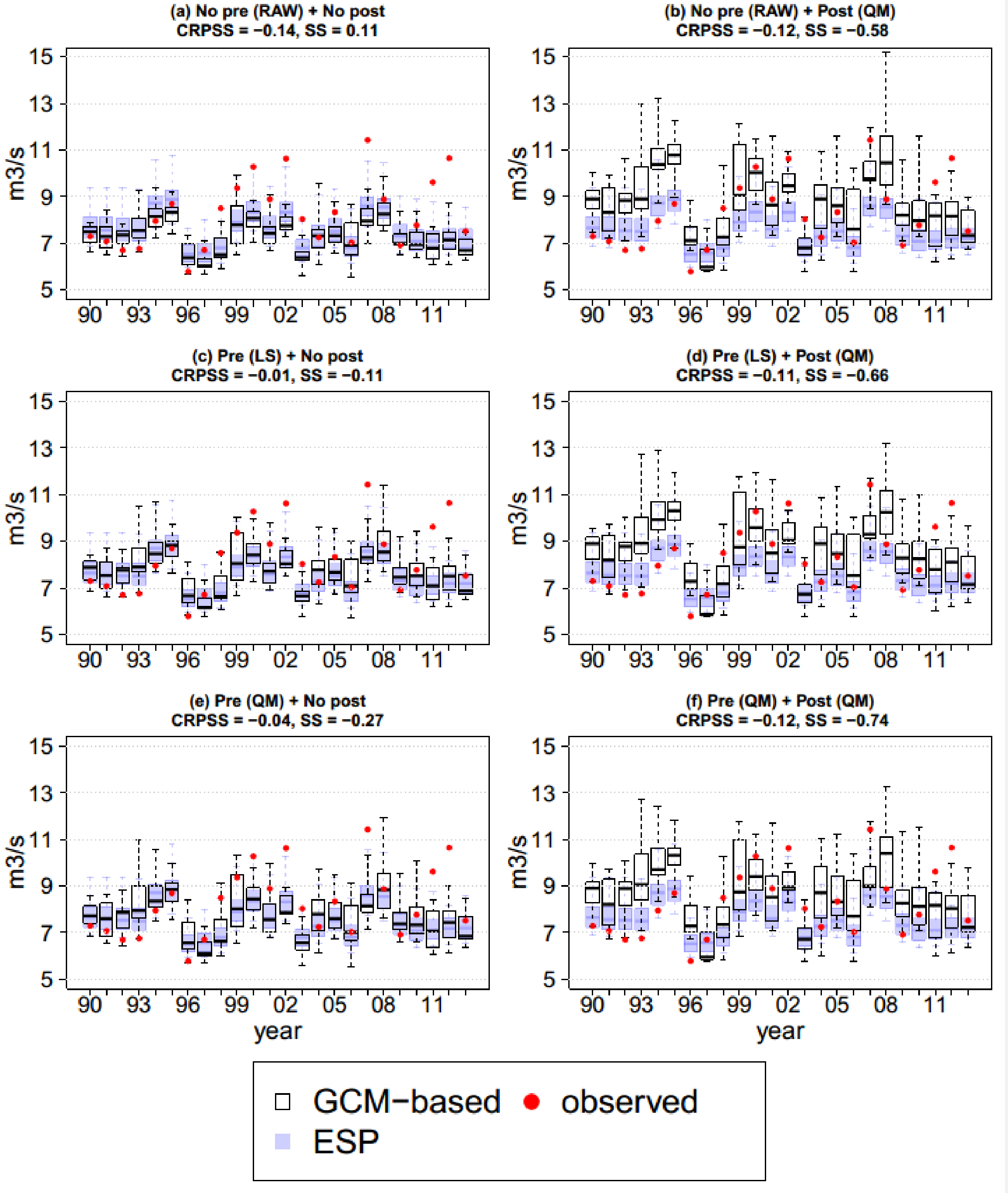
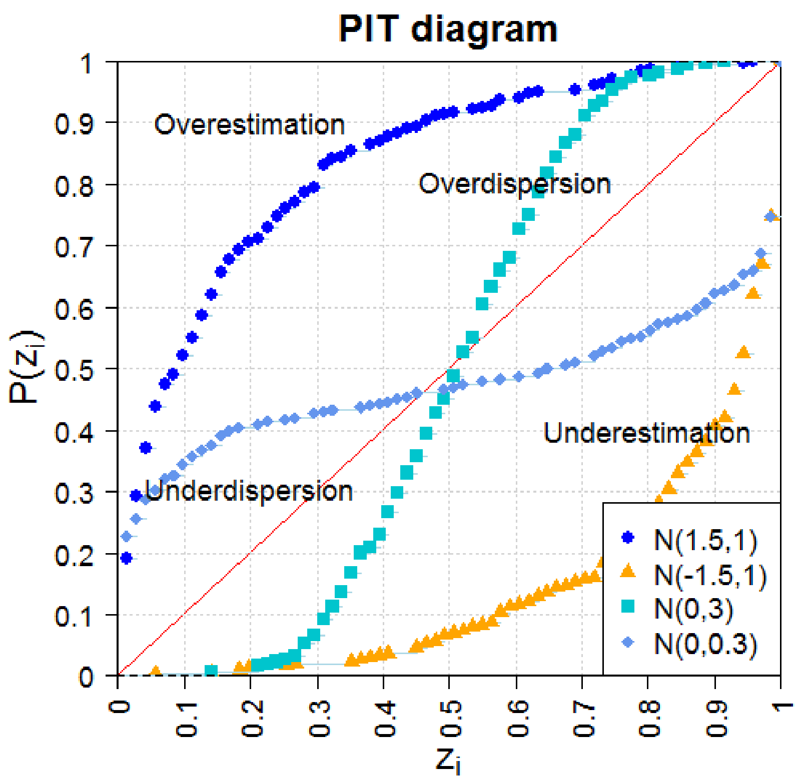


Figure 11: Low flow forecasts for the years 1990-2013 for generated using raw forcings (a), preprocessed forcings with the LS (c) and the QM (e) method and postprocessed streamflow for forecasts generated with raw (b) and preprocessed inputs (d and f). The forecasts are initiated in April. Blue shaded box-plots are ESP forecasts. CRPSS and SS are computed using Eq. (4) with ESP as reference.



~~**Figure A1:** Interpretation of the Probabilistic Integral Transform Diagram (PIT). Observed values are generated with a standard normal distribution $N(0,1)$. Instances of bias in both mean and dispersion are generated with the following distributions $N(1.5,1)$, $N(-1.5,1)$, $N(0,3)$ and $N(0,0.3)$ for an overestimated, underestimated, overdispersive and underdispersive system, respectively.~~

Formatted: Font: (Default) Times New Roman, 10 pt, Not Italic