Reviewer #2:

General comment

R. In general the paper is well written. However, the results section is a bit lengthy (large number of figures and tables). I think it can be condensed without loosing any information. Further, the manuscript would highly benefit from showing some standard verification metrics like the CRPS (Matheson and Winkler 1976; Gneiting and Raftery 2007) and histograms of probability integral transforms (PIT; Diebold et al. 1998, Gneiting et al. 2007).

A. Thanks for pointing out those issues. In the manuscript we removed one table (originally Table 2) that contained redundancies with other tables/figures and moved Tables 3 and 4 (now A1 and A2) to an Annex. Therefore, the main body of the new manuscript only includes three tables, which improve the readability of the paper. As requested, we added analyses of the continuous rank probability score (CRPS). The tables 2, 3, A1 and A2 (this two last tables are now in the Annex section) now include this metric. As in the study we verified the forecasts based on a discrete ensemble (for comparison purposes with the Bias correction forecasts) we incorporate the rank histogram, rather than the histogram of probability integral transforms (PIT). The rank histogram is the analogous tool to the PIT for ensemble forecasts. As you suggested we added the histograms as subpanels in Fig. 3 and 7.

Specific comments

R. As mentioned by reviewer 1 the post-processing methods NGR, BMA, and AKD are not novel and have been used for over ten years in the field of probabilistic (hydro-)meteorological forecasting. Hence, I highly recommend not to use words like "novel" so many times, even though to my knowledge these methods haven't been applied to ETo so far.

We agree that we used inaccurate terms to qualify the methods. This has been amended in the new manuscript. Only in one occasion we refer to "the use of new ETo forecasting strategies" to emphasize that the strategies are for the first time applied to ETo, as you have pointed out. In a few cases we now also use the term "state of art" to qualify the methods, which is commonly managed in literature (for example Gneiting, 2014).

References

Gneiting, T.: Calibration of medium-range weather forecasts, European Centre for Medium-Range Weather Forecasts, Technical Memorandum No. 71, 2014.

R. Reviewer 1 mentioned that the continuous ranked probability score (CRPS) should be included for the verification of the probabilistic forecasts. I strongly agree on this, as the CRPS is a widely accepted verification score that considers errors in location and spread simultaneously.

As stated above we have included this metric. Notice that section 2.4 we now provide the equations and detailed definition of variables and parameters therein of each of the performance metrics used.

R. Further, PIT histograms of the different forecasts should be shown. Like the spread skill plots they measure reliability. However, PIT histograms may help to detect additional issues like potentially wrongly specified parametric distributions in NGR. Subpanels with PIT histograms for selected forecasts could be added to Figures 3 and 7.

A. Thanks for pointing this out. Please see above response to the general comments.

R. The figures captions are not detailed enough. I recommend to move some of the figure descriptions from the main text to the figure captions.

A. We have amended this as requested. Specifically, we modified captions for Fig. 2, 3, 4, 6 and 7.

R. Tables 3 and 6 show only positive biases even after statistical post-processing. However, for a well calibrated forecast, I would expect that the bias is zero in expectation with some random fluctuations. Accordingly, I would expect to see some negative biases for the post-processed forecasts in tables 3 and 6. Why is this not the case?

The post-processing methods tend to slightly overestimate the daily ETo forecasts, as shown by the rank histograms in Fig. 3. The ME values reported in Table 3 (now Table 2) reflect this issue, which is discussed in the revised manuscript. Scheuerer and Büermann (2014) reported similar issues when post-processing ensemble forecasts of temperatures with the NGR method and a version of the BMA method. The weekly forecasts were less affected by these issues, as shown by the rank histogram in Fig. 7. The box plot in Figure 6 shows that the biases over the different regions can be both positive and negative. However, the overall ME (involving all regions) is positive, as reported in Table 6 (now Table 3).

Reference

Scheuerer, M. and Büermann, L., 2014. Spatially adaptive post‐processing of ensemble forecasts for temperature. Journal of the Royal Statistical Society: Series C (Applied Statistics), 63(3), pp.405-422.

R. - L42-45: Please rephrase in a way that it becomes clear that the methods developed by Gneiting et al. (2005) and Raftery et al. (2005) correct also errors in the variance.

A. We agree that the phrase can cause confusion. Now we say:
"Post-processing methods …, are highly recommended to attenuate, or even eliminate, those inconsistencies (Wilks, 2006). Until a few years ago, most post-processing applications only considered single-model predictions …"

Note that we replaced those references by a new reference (Wilks, 2006), which could also help to clarify the idea.

Reference

Wilks, D.S.: Comparison of ensemble-MOS methods in the Lorenz'96 setting, Meteorological Applications, 13(3), pp.243-256, 2006.

R. L66-68: Are the presented post-processing approaches really suitable for extreme values? Being some kind of extended regression models I expect the post-processing methods to perform well in most of the forecast cases, but not for extremes, for which appropriate training data are typically very rare. Further, for extremes the skill of the post-processing approaches highly depends on the shape of the parametric distributions used. As the post-processing approaches presented in this manuscript are all based on normal distributions, there is no possibility to specifically fit the tails of the forecast distributions, which are key when it comes to extremes. Accordingly, I recommend to either remove the sentence on lines 65-66 or explain your reasoning in more detail.

A. Willian et al (2014) compared several methods and found that the methods tested in our study performed well for predicting extreme events. Note the while the AKD and BMA use a Gaussian "dressing" for the ensemble members, the resulting pdf is not necessarily Gaussian. In this sense these methods are more flexible than the NGR method. It may happen (at least in theory) that even if the extreme event is not represented in the training data it can be "detected" by the post-processing method based on the anomalous pattern of the current ensemble. The new sentence is now as:
"The appropriate representation of the second and higher moments of the ETo forecast probability density is especially important to predict extreme values, as shown by Williams et al. (2014)."

Reference
Williams, R. M., Ferro, C. A. T. and Kwasniok, F.: A comparison of ensemble post-processing methods for extreme events. Quarterly Journal of the Royal Meteorological Society, 140(680), pp.1112-1120, 2014.

R. L76: "AF : : : rely on the mean of retrospective reforecasts, thus neglecting information about their dispersion" This is not generally true for analog forecasts. Using a similarity measure that considers also the second moment of the forecast distribution may allow to consider dispersion as well.

A. Thank you for pointing out this. In the modified sentence we excluded the AF method.

R. - L93: ": : :while better fits the user's actual needs" This part of the sentence is a bit out of sync and difficult to understand. What do you mean?

A. Thank you for pointing this out. Yes, we agree. We have removed this part. Note that this paragraph has now been included in the first paragraph of the Introduction.

R - L116-118: Exchangeability is not relevant here, because only ensemble statistics are considered.

A. Yes, this is correct. Thank you for pointing this out. We have removed this term.

- L128: I assume that z denotes a vector? Does u(z) then map the z to a scalar?

A. Yes *u(z) is the variance of z.*

R. - L346-L347: PIT histograms would help to corroborate this

A. Thank you for pointing this out. We have clarified this in the revised manuscript.

R. Technical corrections
- Equations: colon prior to equations can be omitted in most of the cases

A. Thank you for noting this. We have removed most of them in the revised manuscript.

R.
- L43: Gneiting instead of Gneitting
- L60: : : :ETo forecasts is considerably: : :
- L71: : : :post-processing of ETo: : :
- L155: : : :generated at a specific: : :
- L156-157: : : :independent from each other.
- L231: : : :the coverage rations in Table 2 provides: : :

A. Thanks. These issues have been addressed as requested.

- Tables 2,4,5,6: "best" values in bold font like in table 3

A. We have modified Table 6 (now Table 3) in the revised manuscript. In 4 and 5 (now A1 and A2) the "best" values are not so evident.

- L270: : : :the weights of the NCEP model: : :
- L330: see L43
- Figure 2: What is the unit of the lead time?
- Figure 2 / caption: Please mention here that results for daily forecast s are shown here. This comment applies also to the other figures: Please indicate whether they show results for daily or weekly forecasts.

A. Thanks. They have been revised as requested.

- Figure 4a: The uppermost line probably shows Brier scores (BS) and not Brier skill scores (BSS). Please modify the caption accordingly.

A. We show the BSS. Note that the regions with lower skill, as the SE, are those with the larger gains in skill with the probabilistic methods,

- Figure 6: Which quantiles are shown by the box-plots?

A. We have clarified this. It accounts for with the 2.5th, 25th, 50th, 75th and 97.5th percentiles of the distribution.

Additional response

In the revised manuscript, we have added a new section in "Discussion" to discuss some of the earlier published results of post-processing ensemble forecasts of temperature, wind speed, and radiation, and how using these post-processed products, instead of the raw forecasts, to construct ETo forecasts would compare to the post-processed ensemble forecasts of ETo of this research.

All the following references have been added in the revised manuscript:

Bremnes, J. B.: Probabilistic Wind Power Forecasts Using Local Quantile Regression, Wind Energy, 7, 47–54, 2004.

Davò, F., Alessandrini, S., Sperati, S., Delle Monache, L., Airoldi, D. and Vespucci, M. T.: Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting, Solar Energy, 134, pp.327-338, 2016

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B. and Searight, K.: Probabilistic weather prediction with an analog ensemble, Monthly Weather Review, 141(10), pp.3498-3516, 2013.

Hagedorn, R., Hamill, T. M. and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. Monthly Weather Review, 136, 2608–2619, 2008.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15(5), pp.559-570, 2000.

Kang, T. H., Kim, Y. O. and Hong, I. P.: Comparison of pre‐and post‐processors for ensemble streamflow prediction, Atmospheric Science Letters, 11(2), pp.153-159, 2010.

Kann, A., Haiden, T. and Wittmann, C.: Combining 2-m temperature nowcasting and short-range ensemble forecasting, Nonlinear Processes in Geophysics, 18, 903–910, 2011.

Kann, A., Wittmann, C., Wang, Y. and Ma, X.: Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. Monthly Weather Review, 137, 3373–3387, 2009. Hagedorn, R.: Using the ECMWF reforecast data set to calibrate EPS forecasts. ECMWF Newsletter, 117, 8–13, 2008.

Madadgar, S., Moradkhani, H. and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles. Hydrological Processes, 28(1), pp.104-122, 2014.

Messner, J. W., Mayr, G. J., Zeileis, A. and Wilks, D. S.: Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance. Mon. Wea. Rev., 142, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1, 2014.

Møller, J. K., Nielsen, H. A., and Madsen, H.: Time-Adaptive Quantile Regression, Computational Statistics & Data Analysis, 52, 1292–1303, 2008.

Osnabrugge, B. V., Uijlenhoet, R. and Weerts, A.: Contribution of potential evaporation forecasts to 10-day streamflow forecast skill for the Rhine River. Hydrology and Earth System Sciences, 23(3), pp.1453-1467, 2019.

Pinson, P., and Madsen, H.: Ensemble-Based Probabilistic Forecasting at Horns Rev, Wind Energy, 12, 137–155, 2009.

Scheuerer, M. and Büermann, L.: Spatially adaptive post‐processing of ensemble forecasts for temperature, Journal of the Royal Statistical Society: Series C (Applied Statistics), 63(3), pp.405-422, 2014.

Sloughter, J. M., Gneiting, T. and Raftery, A. E.: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging, Journal of the american statistical association, 105(489), pp.25-35, 2010.

Vanvyve, E., Delle Monache, L., Monaghan, A.J. and Pinto, J.O.: Wind resource estimates with an analog ensemble approach, Renewable Energy, 74, pp.761-773, 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, Journal of Hydrology, 501, pp.73-91, 2013.

Verzijlbergh, R. A., Heijnen, P. W., de Roode, S. R., Los, A. and Jonker, H. J.: Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications, Solar Energy, 118, pp.634-645, 2015

Whan, K. and Schmeits, M: Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods. Mon. Wea. Rev., 146, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1., 2018.

Wilks, D. S.: Multivariate ensemble Model Output Statistics using empirical copulas, Quarterly Journal of the Royal Meteorological Society, 141(688), pp.945-952, 2015.

Wilks, D.S.: Comparison of ensemble-MOS methods in the Lorenz'96 setting, Meteorological Applications, 13(3), pp.243-256, 2006.

Wilks, D.S.: Extending logistic regression to provide full probability distribution MOS forecasts, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 16(3), pp.361-368, 2009.

Wilks, D.S.: Sampling distributions of the Brier score and Brier skill score under serial dependence, Q J Roy Meteor Soc, 136(653): 2109-2118, 2010.

Williams, R. M., Ferro, C. A. T. and Kwasniok, F.: A comparison of ensemble post-processing methods for extreme events. Quarterly Journal of the Royal Meteorological Society, 140(680), pp.1112-1120, 2014.

Zhang, J., Draxl, C., Hopson, T., Delle Monache, L., Vanvyve, E. and Hodge, B. M.: Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods, Applied Energy, 156, pp.528-541, 2015.