Interactive discussion on «Subseasonal hydrometeorological ensemble predictions in small- to medium-size mountainous catchments: Benefits of the NWP approach» by Monhart et al.

Anonymous Referee #1

Review on the paper by Samuel Monhart et al. Subseasonal hydrometeorological ensemble predictions in small- and medium-size mountainous catchments: Benefits of the NWP approach Presented for the review is a paper dedicated to the evaluation of subseasonal streamflow forecasts performance in three mountainous catchments in Switzerland produced by the two approaches. Both approaches involve a regional hydrological process-based model PREVAH to account for the initial conditions in the catchment under consideration and the main concern of the paper is concentrated in the model driving for the forecast lead-times. The first approach is the Ensemble Streamflow Prediction (ESP) framework as described by Day (1985), which uses the historical weather data to force the model for the forecast lead-time resulting in the ensemble of the streamflow hydrographs. The second approach is the Numerical Weather Prediction (NWP) framework that involves a meteorological large-domain model-based 5-member forecasts by the ECMWF IFS to force the hydrological model for the lead-time period. The authors use both the raw and bias-corrected NWP forecasts in terms of meteorological and hydrological forecasting skill. The performance of the forecasting approaches is evaluated both for deterministic and probabilistic properties, e.g. the average characteristics are benchmarked by the MAE, NSE and NSElog criteria and the ensemble spread is evaluated by CRPSS metric, as well as the forecast spread to error ratio; the reliability of the forecasts is further examined by constructing the rank histograms. The overall importance of the study is crucial beyond doubt, as is very well described in the Introduction section – the ensemble forecasting methodology is now employed in many forecasting centers around the globe, yet the mentioned improvement in the NWP systems that hydrological prediction systems may benefit from is achieved mainly in Europe and North America, where the outstanding effort to it is applied. The case study catchments choice matches the research aims very well, as very diverse streamflow generation conditions are within the scope of the study – snowmelt-driven and fast-responding catchments are considered, which are an effort in constructing a well-performing streamflow model, as well as reliable subseasonal forecast, especially for summer and fall rainy periods with short hydrological system memory. Still, the authors show good model evaluation metrics. The main findings in the paper are in different effects of NWP bias-correction on the forecast performance, which vary in terms of variable, space and time, e.g. pre-processing of the input forcing is evaluated for temperature and precipitation apart and combined, and discussed for the three catchments over several seasons. The forecasts performance is evaluated not only for the streamflow but for the snow water equivalent in the catchments, as well, which is very crucial for the understanding of the predictability of snowmelt runoff. The results show the NWP pre-processed temperature forecasts outperforming the ESP forecasts, which is a crucial finding, as well.

My overall perception of the paper is that it presents an outstanding scientific effort, which is of critical importance to the modern hydrological forecasting systems research. The motivation is well described, the methods are concise and well referenced, the results are well documented and discussed and a number of very crucial statements on the topic are made. I would recommend minor revisions before the paper can be published.

Specific replies:
Anonymous Referee #1
A few concerns that I would appreciate the authors to enhance in the paper.
First is the statement in section 3.2 concerning the minor importance of such variables as relative humidity etc. on the model performance on such timescales. Hence, the authors state that the relative humidity values were taken from the forecasts without any pre-processing. I would not agree with the authors on the minor importance of the relative humidity, as the evaporation rate is highly dependent on the relative humidity, especially within the processes of evaporation from snow. Given that the temperature forecasts are pre-processed, the close relation of the relative humidity and temperature may influence the forecast performance. However, the assessment of this was beyond the scope of the study.

Reply:
As you mention, variables such as relative humidity play a crucial role regarding the evaporation rates. With our statement we do not intend to undervalue the effect of relative humidity on the hydrological forecast. We rather aim at highlighting that uncertainties at the subseasonal timescale which are already large in case of precipitation and temperature forecasts and their effect on the streamflow has not yet been investigated for the setup presented in this analysis. Hence, we believe that not pre-processing relative humidity (and other variables like sunshine duration) is justified within this analysis but should be further investigated in future studies. However, this will be a challenging task and it might be necessary to choose a different approach, as observational data at the grid scale is at least to date not available in Switzerland.
Hence, we agree with your comments and reformulated the corresponding paragraph by mentioning the importance the additional variables and pointing out that a conclusive analysis of pre-processing additional variables such as relative humidity is be beyond the scope of this study.

To account for this comment, we decide to weaken the statement in the revised manuscript.

We replaced the sentence “In addition, these variables are of minor importance for the forecst time scale investigated in this study” with the following sentence:

P7; L4-6
“Although these parameters could influence the hydrological relevant processes, e.g. evaporation rates from snow based on the relative humidity, a thorough assessment of the effect of bias correcting and downscaling of these additional variables is out of scope of the current study.”

Anonymous Referee #1
In the SWE forecasts verification section 4.3.4 the forecasts are verified against the reference model run instead of the actual observation, yet I would appreciate if the reference model performance could be discussed at least within a few sentences. Another consideration is that the methodology of the SWE assessment should be placed in the corresponding subsection within section 2.

Reply:
We accounted for this by including a brief discussion of the findings presented in Jörg-Hess et al. (2014) where they found good agreement between the predicted SWE and the observed SWE maps for the entire 32 day forecast horizon.

The new sentence read:

P12; L 3-5
“A verification of modelled SWE against a consistent and homogenized climatology of gridded SWE based on station information is given by Jörg-Hess et al. (2014). They have shown that the modelled SWE exhibit errors that are in the same order as natural variability. “

In addition, we now mention in section 2 that the verification is performed for both variables, streamflow and SWE.

P5; L17-18
The verification is performed for the two variables streamflow and snow water equivalent (SWE).

Anonymous Referee #1

Anonymous Referee #1
Minor technical note: p. 13 l. 1 -
“...seasonal meteorological...” must be followed with a noun, which is missing

Reply:
Done: “…seasonal meteorological predictions…”

Anonymous Referee #2

General comments This is an interesting and thorough assessment of an ensemble streamflow forecasting system in snow-affected mountainous catchments. The system pairs NWP forecasts with a distributed hydrological model that includes detailed accounting of cryogenic processes. The system is technically advanced and in my opinion of high interest to the readership of HESS. The study is well conceived and very clearly written. A particular strength of the study is the extensive and thorough verification of the forecasts, encompassing multiple appropriate measures of performance that are described and discussed in clear and interesting ways. In general, the authors' conclusions are strongly supported by their analyses. I have one quibble with the use of QM as a means for downscaling meteorological forecasts, which amounts to a minor revision. Other than this, I have no hesitation in recommending this study for publication.

General comments:

Anonymous Referee #2
It appears that the authors use QM to downscale NWP predictions from a coarse grid (_30/60 k) to a fine grid (_2 k). It's well established that QM is not theoretically ideal for this practice, because of so-called variance inflation (Maraun 2013). The authors appear to be aware of this, as they discuss this issue in an accompanying paper (Monhart et al. 2018). However, it is more salient in this paper, because of the hydrological modelling that is carried out. Variance inflation is only an issue when quantile mapped/downscaled meteorological forecasts are spatially reaggregated, which is exactly what the hydrological model does. So it will not show up in the analyses carried out by Monhart et al. 2018 (where variables are not reaggregated), but it could well be an issue in this study.

In addition, and as Maraun shows, the variance inflation problem is only strongly evident for extremes. Extremes are not the focus of the analysis carried out in this paper, which is fine. But this means it's hard to tell if variance inflation is present in streamflow forecasts. As the forecasting system could be used for flood prediction, this may be a serious issue. Accordingly, I recommend two changes to the manuscript:

1) The authors should clearly describe how they bridge the gap in spatial resolution from a _30/60 k horizontal grid (NWP forecasts) to a _2 k horizontal grid (observations).

Reply:
We now included a more concise description on how the gap between the coarse spatial resolution of the NWP predictions and the 2 km gridded observation is performed. To account for this concern, we expanded subsection 2.1 where the pre-processing step using QM is described.

P4; L10-15
"The pre-processing is performed for temperature and precipitation and involves not only a bias correction but also a downscaling because of the higher resolution of the gridded observation data used in this study. The observation and forecast data used in this study is described in more detail in the Section 3. However, it is worth mentioning here that the raw model resolution of 50 km is bias corrected with QM using gridded observations with a

higher spatial resolution of 2km. This resolution corresponds to the meteorological input of the hydrological model, for which observations from station data are interpolated to 2 km grids (see section 3.2)."

Anonymous Referee #2
And if, as I've assumed, they use QM for this purpose:
2) The authors should briefly acknowledge the issue of variance inflation in the discussion, including a discussion of possible implications for their system (perhaps alongside recommendations for dealing with these implications).

Reply:
As you correctly assume (and as we already answered in the short reply) we do use QM to downscale the predictions and we are aware of the variance inflation issue caused by this method. We now included a new paragraph in the discussion to highlight the limitations of QM related to the variance inflation. We acknowledge that the problem of variance inflation can influence the results and argue why we believe it is still justified to use QM for the downscaling within our study. As you suggested we now include reference that discuss the problem in more detail and propose alternative approaches that could be used in future studies. The added paragraph reads as follows:

P13; L7-L21
"An additional critical limitation of QM is the issue of variance inflation. Maraun (2013) emphasizes that the variance of the downscaled product strongly depends on the variance of the raw model grid box and QM does not introduce any small-scale variability. This is of particular importance for applications using local-scale information (such as distributed hydrological modelling) and if extremes are considered. In CH2018 (2018) these limitations of the QM method are highlighted for local climate change scenarios in Switzerland. In particular for convective precipitation events in summer the variance inflation issue can cause misinterpretation of data at the finer resolved scale. In the present study we are interested in the average streamflow throughout all season in the year for the upcoming 32 days and not in predicting extremes what reduces these implications, but still the spatial structure especially during convective situations in summer will likely be misrepresented and can influence the results. Different alternatives could be used depending on the specific application of the downscaled information, e.g. perfect prognosis approaches (Von Storch, 1999), the use of weather generators (e.g. Peleg et al., 2017) or in general stochastic methods (e.g. Volosciuk et al., 2017). However, such methods often require large computational resources. As the intention of this study is to pioneer the use of subseasonal hydrological predictions towards an operational use, we decided to use the QM technique despite its limitations. The results presented above and discussed in the following paragraphs legitimate our choice. Nevertheless, future studies should focus on the effect of variance inflation when QM is used to pre-process the predictions and alternative methods should be considered."

Anonymous Referee #2
I also encourage the authors to consider measuring the impact of variance inflation on their system in future work.

Reply:
We will consider this issue in our future work by either choosing a different correction technique and/or a careful assessment of the effect resulting from the inflation issue.

Anonymous Referee #2
Specific comments
Page 4 L7-9 "For a given target day of a reforecast the correction is derived from the distribution of all the reforecasts within a three weeks window around the same lead day and the corresponding observations, hence the correction depends both on the lead time and on the period of the years".
Is the QM cross-validated in some way?

Reply:
Yes, the forecasts are cross-validated. The calibration of the hindcasts is performed in a leave-one-year-out cross-calibration framework. Hence in the verification the information of the year to be verified is not used in the calibration procedure i.e. cross-validated. We added the following statement to make that clear.

P4; L8-9
"This cross-calibration framework ensures a cross-validation described in subsection 2.3."

Anonymous Referee #2
How are zero values in precipitation handled in the QM?

Reply:
Zero values are not handled in a specific way. Because we do apply a multiplicative correction zero values will not cause an issue. Therefore, the QM does not artificially produce rain, i.e. in case zero precipitation is forecasted the corrected precipitation still has zero precipitation. In general, weather prediction models exhibit a drizzle effect due to their large grid sizes, meaning that raw model forecast generally predict too much rain compare to the observation (in case of very low precipitation rates). Hence, zero precipitation values do not need special treatment during the pre-processing.
We now included an additional sentence to highlight that the in the multiplicative version of QM zero values do not need special treatment in the pre-processing step.

P 4; L20-21
"Using the multiplicative version of QM for temperature allows to include zero value without special treatment. Hence, no precipitation can be generated if the raw forecasts do not exhibit any rain."

Anonymous Referee #2
Page 5 Section 2.4 For each score used, please note the range of values taken and the orientation of the score (e.g. -infinity to 1 for NSE, with 1 being perfect) This allows easy intepretation of, e.g., Figure 4.

Reply:
We now include the following statement in the figure caption.

P31, L11-13
"The NSE and the CRPSS span from -infinty to 1 with a perfect score being 1; a bias of zero indicates no forecast error with negative values indicating underestimation and positive values indicating overestimation of the flow; reliable forecasts exhibit a SprErr of 1 and lower values indicate overconfidence and greater values indicate overdispersion."

Anonymous Referee #2
L16 "we use the spread to error ratio (SprErr) as an indicator for the forecast reliability" Please briefly describe how this is calculated.

Reply:
We now added a sentence to describe how the spread to error ratio is calculated:

P5; L29-32
"The *SprErr* is defined as the ratio between the variance of the forecast ensemble (forecast spread) and the mean squared error (MSE) of the ensemble forecast (forecast error). For reliable forecasts the spread and the error are equal, resulting in a SprErr of 1 whereas values below 1 indicate overconfidence (errors are larger compared to the spread) and values above 1 indicate overdispersion (the spread is larger compared to the error)."

Anonymous Referee #2
Page 8 L18 Figure 3. It's very difficult to see the different colours in this figure, especially in the right hand panel - i.e., it's not possible to distinguish QM from raw. Please replot so this is clearer (e.g., with different colours/box outlines, and/or perhaps restrict the vertical axis in the rh panel to [-0.2 0.3]).

Reply:
We change the figure according to your suggestions.

Anonymous Referee #2
L22-23 "After bias correction the skill is higher with positive CRPSS up to three weeks in winter and spring." From the figure, precip skill looks to be negligible in DJF after week 1. Skill scores will of course be a little noisy (in time, as well as in space; the authors have only considered spatial variation), so I don't think the authors should describe forecasts as 'skillful' if they have CRPSS values only very slightly above zero.

Reply:
We agree, the description is too optimistic. We considered this comment and only mention spring in this sentence. It now reads:

P9; L19-20
"After bias correction the skill is higher with positive CRPSS up to three weeks in MAM. In JJA the positive skill is only observed for 2 weeks lead time and in SON and DJF for the first week only."

Anonymous Referee #2
L28 NSE, Bias - I assume these are calculated on the mean of the ensemble? Please state this in Section 2.4.

Reply:
Your assumption is correct. This will be included in Section 2.4

Page 5; L23-24:
"For both versions of the NSE and the bias the ensemble mean is used for the calculation. "

Anonymous Referee #2
Page 9
L7-8 "The negative biases of the ESP approach indicate an underestimation of the streamflows for all lead times in the Verzasca catchment." It's not clear to me why ESP predictions would be biased. ESP forcings, by construction, are unbiased. As the bias in predictions is calculated against model climatology, there should be no bias, as occurs in the Thur. Please briefly explain what is going on here.

Reply:
We generally agree with your statement that ESP predictions are by construction unbiased. The historical meteorological observations used to run the hydrological simulations are a sample of the climatology and thus the resulting streamflow prediction should theoretically agree (i.e. be unbiased) with the climatological streamflow used for the verification. However, there are two effects that might lead to biases in the ESP prediction. First, if the initial conditions at the time of the forecast initialization strongly deviates from the climatology the ESP prediction will take more time until the streamflow converges with climatology, in particular in snow-dominated catchments as the Verzasca and the Klöntal catchment. Second, the meteorological input for the ESP predictions in our case is not an exact sample of the climatology. The meteorological observations from 1980 to 2014 are used, the streamflow climatology is based on the period 1994-2015. Hence, trends in the meteorological input might affect the streamflow prediction resulting in a bias. E.g. temperature exhibits a positive trend within this period, colder temperature in the beginning of the period might lead to negative biases (underestimation of the streamflow) because the storage of precipitation in snow is enhanced. Again, this effect is stronger in snow-dominated catchments.

Anonymous Referee #2
L15 "The spread error ratio of the ESP predictions is below 1 for all lead times indicating overconfidence." There are two issues here. First, as I already alluded to, it would be beneficial to readers unfamiliar with the spread-error ratio to offer a brief explanation of the range of values it can take, and which direction indicates over/under confidence in Section 2.4.

Reply:
We now adapted the description in the methodology.

P5, L30-32:
"For reliable forecasts the spread and the error are equal, resulting in a SprErr of 1 whereas values below 1 indicate overconfidence (errors are larger compared to the spread) and values above 1 indicate overdispersion (the spread is larger compared to the error)."

Anonymous Referee #2
Second, I can't understand why the ESP forecasts are not reliable. ESP forcings are by construction reliable, so the spread-error ratio for the ESP forecasts should be close to 1. But this is not so in the two smaller catchments. Why is this?

Reply:
We suggest a similar explanation as in the answer to the comment related to the bias of the ESP (first answer on this page). In short, different time period of the meteorological forcing

of the ESP (1980-2014) and the streamflow climatology (1994-2015) might affect the SprErr as well. The forecast in the longer period might lead to larger error what could lead to a reduced SprErr. In addition, the effect of initial conditions that are strongly deviating from the climatological state can lead to larger errors especially at early lead times until the predictions converge to the climatology.

Anonymous Referee #2
Page 10
L26-27 "The rank histograms for the ESP predictions do provide more uniform rank histograms with a weak tendency of a negative bias." Again, I would be interested in a brief explanation of this bias in ESP forecasts.

Reply:
See answer to the comment above.

Anonymous Referee #2
Page 12
L15 "QM indeed is able to provide reliable ensembles" To me this sounds as though QM is responsible for the reliable ensembles, and this isn't really correct. QM can improve reliability to the extent that it is negatively impacted by bias (including conditional bias - i.e., biases at different points in the marginal distribution). You have shown (like Zhao et al.) that QM does not correct for overconfidence in short lead-time forecasts. This is because the underlying forecasts are overconfident - QM can't correct this (by construction). At longer lead-times, the QM forecasts are reliable because the spread in the underlying forecasts is appropriate (notwithstanding conditional biases). The same goes for coherence (discussed in the accompanying Monhart et al. 2018 JGR paper evaluating the NWP forecasts) - QM is not capable of correcting negative skill wrt to climatology in forecasts, other than that due to bias (again, by construction). The forecasts presented in this study are coherent because the underlying (raw) forecasts appear to be neutrally skillful at long time scales. In other words, it is the combination of NWP model and the QM that create the reliable and coherent forecasts shown in this study, not just QM. I think it would be better to reword what's written here to reflect this fact.

Reply:
We account for this comment and reworded the statement to make it clear that QM alone is not able to provide reliable forecasts if the underlying meteorological forecast are not reliable.
The sentence now reads:

P13; L4-7
"To summarize, it was found that the combination of the NWP model with QM indeed is able to provide reliable ensembles for lead times beyond 10 days but at shorter lead times the ensembles tend to be overconfident because the spread in the underlying NWP forecasts tends to be inappropriate what cannot be corrected using QM."

Anonymous Referee #2
P14
L9-10 "to correct the errors and biases of the hydrological simulation" and, presumably, to account for additional uncertainty induced by the hydrological model in the ensemble?

Reply:
Thanks for this suggestion. We included this, and the sentence now reads:

P15; L32-33
"…to correct the errors and biases of the hydrological simulation and to account for additional uncertainty induced by the hydrological model in the ensemble."

Anonymous Referee #2
P14 "To do so we verified the streamflows of the reference simulation of the PREVAH model against observations." One thing not discussed here is reliability. I assume when assessed against observations, the ensembles are highly overconfident because uncertainty in the hydrological model is not included in the ensemble (see, e.g., Bennett et al. 2014). This is especially true at very short lead times (perhaps <3 days), when hydrological model uncertainty may be the dominant source of uncertainty in the forecasts. This may be worth mentioning.

Reply:
Here we discuss the performance of the reference simulation verified against observations. The reference simulation is a single timeseries and thus no ensembles can be verified. But for the ensemble predictions verified against the observation we indeed find a pronounced overconfidence at short lead times. Hence the results confirm what you mention in your comment. Although we decide not to include any additional figures we mention this aspect in the discussion.

P16, L13-L18:
"An example of such a deficiency is the uncertainty resulting from the hydrological modelling that result in stronger overconfidence especially at short lead times when the hydrological model uncertainty may be the dominant source of uncertainty as discussed for example in Bennett et al., (2014). To illustrate this for the prediction used in this study, the verification was repeated with the real runoff observations. The skill in most seasons (MAM, JJA, SON) exhibit the same behavior with high skill at early lead times and decreasing skill at longer lead times and generally higher overconfidence at short lead times is observed what confirms the findings by Bennett et al. (2014)."

Typos/Grammar Page
1 L14 "Prior of" should be "Prior to" or more simply "Before"
Done

Page 2 L11 "both," delete comma
Done

Page 5
L2 "comparison of to" delete "of"
Done

L28 "year" should be "years"
Done

Page 7 L3 "station" should be "stations"

Done

Page 8 L9 "evaluate of operational" delete "of"
Done

Page 9 L9 "and reach" should be "and reaches"
Done

L11 "enhance" should be "enhances"
Done

L11 "elongates positive up" I think "skill" is missing here - i.e. "elongates positive skill up"
Done

L27 "are shown" should be "is shown"
Done

Page 13 L1 "in seasonal meteorological can" I think this should be "in seasonal meteorological forecasts can"
Done

Page 14 L11 "prediction" should be "predictions"
Done

L23 "exhibit" should be "exhibits"
Done

Page 15 L14 "enhance" should be "enhances"
Done

L24 "EPS" should be "ESP"
Done

L25 "both, temperature" delete comma
Done

Page 27 L5 "whereas" should be "where"
Done

References Bennett JC, Robertson DE, Shrestha DL, Wang QJ, Enever D, Hapuarachchi P, Tuteja NK. 2014. A system for continuous hydrological ensemble forecasting (SCHEF) to lead times of 9 days. Journal of Hydrology 519: 2832-2846. DOI: 10.1016/j.jhydrol.2014.08.010.
Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C. and Liniger, M. A.: Skill of Sub-seasonal Forecasts in Europe: Effect of Bias Correction and Downscaling using Surface Observations, J. Geophys. Res. Atmos., 1–18, doi:10.1029/2017JD027923, 2018.

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E. and Ramos, M. H.: How suitable is quantile mapping for postprocessing GCM precipitation forecasts?, J. Clim., 30(9), 3185–3196, doi:10.1175/JCLI-D-16-0652.1, 2017.
Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-458, 2018.

Anonymous Referee #3
General comments.

The main motivation of this study is to fill the gap in small scale researches in determination of the propagation positive skill extent in meteorological prediction models further into the streamflow forecasts. To address this problem, a traditional
ESP approach was compared with prediction driven by ECMWF subseasonal ensemble system in three alpine catchments with varying hydroclimatic conditions. To emphasis the effect of applying pre-processing (QM-based) of NWP output, prediction verification was done against the reference simulation (pseudo observations). Thus hydrological model errors were excluded from the analysis.
Summary.
There was in-depth discussion on hydroclimatic variability and predictability, the role of forcing and model parameters' uncertainty. The verification metrics used were relevant and applied in a logical manner. The results well supported the conclusions. Some sections recommend recompiling for better logically organized and easy follow.

I recommend publishing the manuscript but encourage the authors:

1) Give a more justification on choice of these watersheds for sub-seasonal forecasting. Initially, it can be supposed that the study is a part of a large numerical experiment where the results were confirmed only for the three arbitrary watersheds.

Reply:
The choice of the watersheds was driven by the application of the forecast and the diversity in a hydroclimatic sense. Within the larger umbrella project, we aim at analysing and quantifying the benefit of using subseasonal forecasts for the optimization of hydropower operations to increase their revenues in Switzerland. Therefore, watersheds with installed hydropower operations were selected first (the Klöntal and the Verzasca catchment, representing snow dominated and partially glaciated catchments). To broaden the scientific value of our analysis and better quantify the effect of snow, we included as well the Thur catchment which is precipitation dominated. Hence, the watersheds are not an arbitrary choice from a large numerical experiment but rather a reasonable compromise between further use of the results for our intendent application and a scientific in-depth analysis of the forecast performance considering different hydroclimatic regimes.
We account for this by explaining the motivation for the choice of the watersheds in section 3.3 of the analysis.

P7; L10-15
"The selection of the catchments is a compromise between the intended applications of our results within hydropower optimization in the Alpine region and to adequately meet the requirements of a scientific analysis. Hence, the Klöntal and the Verzasca catchment both are selected because of existing hydropower installation in these watersheds and the Thur catchment was chosen as a representative catchment with different hydroclimatic characteristics and because the catchment is often considered in hydrological research in Switzerland."

Anonymous Referee #3
2) Add meteorological observations network on fig. 2 and give some comments explaining the good modeling quality (tab. 1) when using the grid product obtained at a low observations network density.

Reply:
As mentioned in the manuscript, the grid product is often used in climate related research in Switzerland. The relevant publication on the gridded product, including verification studies of the gridded product are given in the manuscript. An overlay of the station used to produce the gridded data set is in our opinion an over overkill for the present study as the information can be found in in the given references. Therefore, we decide not to include the observational stations in figure 2. The good modelling quality, speaking of the performance of the streamflow predictions indicate that the gridded product does provide a good baseline for pre-processing subseasonal hydrological predictions.

We included the following sentence in the discussion:
P14, L1-3:
"Furthermore, the improvements in performance of the streamflow predictions by pre-processing suggests that the gridded observational dataset provide a good baseline for this purpose, despite the difficulties involved in producing a gridded product based on a limited number of observational stations."

Anonymous Referee #3
3) Give a number of predictions made for evaluation.

Reply:
For the verification we use a total number of 1040 reforecasts with 5 members each. This number results from the dataset which provides 1 reforecast per week for the 20-year period (1994-2014). For the seasonal aggregation a total of 260 reforecasts is used for the verification.

We added this in the section 3.1. the sentence now reads:

P6; L20-21
"… covering the period from April 1994 to March 2014 resulting in a total of 1040 individual reforecasts that are analyzed within this study."

Anonymous Referee #3
4) Comment on how the processing of only temperature and precipitation affects and propagate through the hydrological simulation.

Reply:
The results clearly show that pre-processing both temperature and precipitation is important to provide well performing hydrological simulation. Hence, pre-processing temperature only does have a large impact on the performance of the streamflow predictions but if precipitation is pre-processed as well, the rank histograms indicate more reliable forecasts. This subject is addressed in the results section 4.3.

Anonymous Referee #3
5) Specify if the ESP method can outperform the NWP if ensemble takes not all but only individual years guided by a certain criterion for the similarity of the initial conditions.

Reply:
As you mention, the performance of the ESP predictions could be enhanced based on a selection of years where the initial conditions show high similarity with present conditions. E.g. Crochemore et al. (2017) showed that seasonal forecasts can benefit by conditioning climatology. However, a thorough assessment of conditioning the ESP predictions is beyond the scope of the present study but could be considered for further studies within this area.

We widened the discussion to make the reader aware of this potential.

P14, L4-7:
"In addition, the performance of the ESP predictions could potentially be enhanced if not all, but only individual years are taken into account. A certain guidance based on a selection of years with similar initial conditions could be taken into account. Crochemore et al. (2017) have shown that seasonal prediction based on ESP can benefit from condition the forecasts on climatology. However, an evaluation of such an approach is out of scope of the present study."

Anonymous Referee #3
6) Recompile the sections 2 and 3 referred to each other to make them more consistent.

Reply:
As a result of including the comments of the all three reviewers, some changes have been made to section 2 and 3 as well. Both sections should now be more consistent.

Specific comments.
P.13, L5. The upper (reads like nested) Thur subcatchment Halden (1750 km2) is little bit bigger then Thur watershed itself (1696 km2).

Reply:
You are right, the Thur subcatchment Halden is 1085 km2. We changed this accordingly.

Subseasonal hydrometeorological ensemble predictions in small- and medium-size mountainous catchments: Benefits of the NWP approach

Samuel Monhart[1,2,3], Massimiliano Zappa[1], Christoph Spirig[2], Christoph Schär[3], and Konrad Bogner[1]

[1]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Mountain Hydrology and Mass Movements, Birmensdorf, Switzerland
[2]Federal Office of Meteorology and Climatology MeteoSwiss, Climate Prediction, Zurich-Airport, Switzerland
[3]ETH Zurich, Institute for Atmospheric and Climate Science, Zurich, Switzerland

Correspondence to: Samuel Monhart (Samuel.Monhart@wsl.ch)

Abstract. Traditional Ensemble Streamflow Prediction systems (ESP) are known to provide a valuable baseline to predict streamflows at the subseasonal to seasonal timescale. They exploit a combination of initial conditions and past meteorological observations, and can often provide useful forecasts of the expected streamflow in the upcoming month. In recent years, numerical weather prediction (NWP) models for subseasonal to seasonal timescales have made large progress and can provide added value to such a traditional ESP approach. ~~Prior of~~Before using such meteorological predictions two major problems need to be solved: the correction of biases, and downscaling to account to increase the spatial resolution. Various methods exist to overcome these problems, but the potential of using NWP information and the relative merit of the different statistical and modeling steps remains open. To address this question, we compare a traditional ESP system with a subseasonal hydrometeorological ensemble prediction system in three alpine catchments with varying hydroclimatic conditions with areas between 80 and 1700 km^2. Uncorrected and corrected (pre-processed) temperature and precipitation reforecasts from the ECMWF subseasonal NWP model are used to run the hydrological simulations and the performance of the resulting streamflow predictions is assessed with commonly used verification scores characterizing different aspects of the forecasts (ensemble mean and spread). Our results indicate that the NWP based approach can provide superior prediction than the ESP approach, especially at shorter lead times. In snow-dominated catchments the pre-processing of the meteorological input further improves the performance of the predictions. This is most pronounced in late winter and spring when snow melting occurs. Moreover, our results highlight the importance of snow related processes for subseasonal streamflow predictions in mountainous regions.

1 Introduction

Subseasonal hydrometeorological predictions are of special interest for many different applications in the public and the private sectors. For example to develop early warning systems for flood and drought preparedness for the general public

(Bogner et al., 2018; Cloke and Pappenberger, 2009; Fundel et al., 2013a; Van Lanen et al., 2016; Schär et al., 2004; White et al., 2017), to optimize the production of renewable energy sources such as wind (Beerli et al., 2017), solar (Inman et al., 2013) and hydropower (García-Morales and Dubus, 2007) or to ensure inland waterway transportation (Meißner et al., 2017).

In recent years the numerical weather prediction (NWP) systems have greatly evolved (Bauer et al., 2015). The ensemble forecasting approach introduced two decades ago allows capturing the chaotic nature of the atmosphere in a probabilistic sense and has extended the horizon to which weather predictions can provide valuable information. Hydrological prediction systems, especially beyond the short-range lead times, benefit from such an ensemble approach (Demargne et al., 2014; Jaun et al., 2008; Schaake et al., 2007; Verbunt et al., 2007). Ensemble hydrometeorological end-to-end prediction systems for the subseasonal time scale, i.e. forecasts for lead times up to 4 to 6 weeks, are now being developed and investigated for different parts in the world. Nowadays, different research initiatives (Hao et al., 2018; Robertson et al., 2015; Vitart et al., 2017; Vitart and Robertson, 2018) set their focus on the assessment of predictability within this lead-time horizon. But still this lead time between the medium-range forecasts and the seasonal predictions is a grey zone in both, the meteorological and hydrological forecasting communities.

However, ensemble prediction systems have extensively been used for short to medium-range forecasts and it could be shown the resolution of the model plays an important role for a good performance of ensemble precipitation forecasts (Marsigli et al., 2008; Montani et al., 2001, 2003). A higher resolution can be achieved by dynamically downscaling the ensemble predictions, what led to a gain in predictability in ensemble precipitation forecast over the past years in the medium-range time scale (Montani et al., 2011). From a hydrological perspective flood peaks can much better be predicted using mesoscale ensemble forecast models compared to global models (Davolio et al., 2012). In such operational short to medium-range forecasting systems, dynamical downscaling methods are a valuable tool to increase the resolution of the forecasts but for long term predictions the computational demand of dynamical downscaling methods is a limiting factor and it has been shown that similar results can be achieved with statistical downscaling methods (Díez et al., 2005; Manzanas et al., 2018). In recent years, seasonal ensemble prediction systems have increasingly been used in the hydrological forecasting context. At this time scale downscaling is usually included in statistical processing techniques that primarily aim at correcting the bias of the meteorological forecast. Different statistical bias correction techniques can be used to improve the skill of seasonal forecasts (Bohn et al., 2010; Crochemore et al., 2016; Kumar et al., 2014). An alternative approach for seasonal streamflow prediction is the traditional ensemble prediction system (ESP) first introduced by Day (1985). In this approach, observed historical meteorological conditions are used to generate the streamflow forecasts. Several studies have shown that Ensemble Streamflow prediction (ESP) can provide skillful seasonal streamflow prediction (Harrigan et al., 2017) and seasonal hydrometeorological prediction systems have difficulties to outperform the ESP beyond one month lead time (Arnal et al., 2018; Lucatero et al., 2018). Irrespective of the methodology used, the performance of ensemble streamflow forecasts depend to a large degree on the initial condition within the catchment especially for small catchments (Van Dijk et al., 2013; Thirel et al., 2010). In subseasonal to seasonal hydrometeorological predictions, the memory of the

river networks is predominately driven by the initial conditions of the land surface characteristics, i.e. soil moisture and snow cover (Jörg-Hess et al., 2015a). It has been shown that these parameters play an important role for skillful hydrometeorological forecasts using numerical weather prediction (NWP) inputs in hydrological models for streamflow forecasting (Orth and Seneviratne, 2013b; Sinha and Sankarasubramanian, 2013). However, for small snow dominated catchments the benefit of using pre-processed subseasonal NWP input has not yet been investigated. The three main reasons for the lack of studies at smaller scales is the fact that the driving meteorological models are calculated on a global scale with coarse resolution. Therefore, small catchments are often smaller than the nominal resolution of these models, making a pre-processing step necessary to bias correct and downscale the meteorological forecasts to an adequate spatial resolution as an input to the hydrological models. The second reason is that small catchments usually do not have a long temporal memory and processes leading to ~~streamflows~~ runoff are rather quick and therefore the skill in early lead times do not extend into longer lead times (Orth and Seneviratne, 2013a). As a consequence, the upper limits of the forecast skill at the subseasonal time scale strongly depend on such catchment characteristics (Bogner et al., 2016, 2018). Finally, previous generations of subseasonal to seasonal climate forecasts rapidly lost skill beyond the first 2 weeks (see for example Lavers et al., 2009).

Small and medium sized catchments in mountainous regions such as the Alps are often snow dominated, hence the streamflow, and especially the forecasts thereof, depends to a large degree to snowmelt processes that are driven by temperature (Hock, 2003; Ohmura, 2001; Zappa et al., 2003). Monhart et al. (2018) have shown that the statistically corrected subseasonal temperature forecasts exhibit positive skill compared to the climatological reference of up to 3 three weeks, whereas positive skill of precipitation forecasts is restricted to the first 10 days. Hence, the question arises if and to what extent the positive skill of temperature forecasts in meteorological prediction models does further progress to the streamflow forecasts. Skillful streamflow forecasts might not directly be related to the skill in temperature forecasts but rather in an appropriate sampling of the initial conditions and the actual snow cover (snow water equivalent) at initialization (Jörg-Hess et al., 2015a). Therefore, the skill of the streamflow forecasts will not solely depend on the quality of temperature forecasts alone. To investigate this question, ensemble streamflow prediction can be used to evaluate the importance of the initial conditions. The comparison of the skill of an ESP forecast and the skill of actual hydrometeorological predictions can indicate the relative importance of using physically consistent meteorological forecasts to produce skillful streamflow predictions. After presenting the methods used for the pre-processing and the verification of the different forecasts in section 2, the meteorological and hydrological data is described in section 3. The results obtained are presented in detail in section 4 followed by a discussion (section 5) to put the results in the current scientific context.

2. Methods

The conceptual framework used in this study is presented in Figure 1. The hydrological model is run with different meteorological forcings to provide ensemble streamflow predictions. A traditional ESP approach using 34 years of

meteorological observations provides the baseline forecasts. A second input is based on the reforecasts from the ECMWF subseasonal prediction system as described in section 3.1. Along this chain four different configurations are used to feed the hydrological model and generate streamflow predictions: the raw reforecasts for both temperature and precipitation, the raw precipitation reforecast and pre-processed temperature reforecasts and vice versa, and both parameters pre-processed.

2.1 Pre-processing in the hydrometeorological model chain

In the pre-processing step a quantile mapping (QM) technique is applied in a leave-one-year-out setup to correct each re-forecast year separately. The corrections are applied in a lead-time dependent manner. A description of the pre-processing setup can be found in Monhart et al. (2018). This cross-calibration framework ensures a cross-validation described in subsection 2.3. In contrast to the site-specific pre-processing setup ~~in that study~~used in Monhart et al. (2018), a gridded observational product is used here. The pre-processing is performed for temperature and precipitation and involves not only a bias correction but also a downscaling because of the higher resolution of the gridded observation data used in this study. The observation and forecast data used in this study is described in more detail in the Section 3. However, it is worth mentioning here that the raw model resolution of 50 km is bias corrected with QM using gridded observations with a higher spatial resolution of 2km. This resolution corresponds to the meteorological input of the hydrological model, for which observations from station data are interpolated to 2 km grids (see section 3.2). The QM technique is a simple and widely used method for pre-processing hydrometeorological forecasts (e.g. Kang et al., 2010; Lucatero et al., 2018; Verkade et al., 2013). For a given target day of a reforecast the correction is derived from the distribution of all the reforecasts within a three weeks window around the same lead day and the corresponding observations, hence the correction depends both on the lead time and on the period of the years. For temperature an additive correction and for precipitation a multiplicative correction is applied. Using the multiplicative version of QM for temperature allows to include zero value without special treatment. Hence, no precipitation can be generated if the raw forecasts do not exhibit any rain.

2.2 Ensemble Streamflow Prediction (ESP)

The Ensemble streamflow predictions follow the established procedure first proposed by Day (1985). Many studies have shown the potential of this method to provide skillful streamflow predictions at the subseasonal to seasonal time scale in Europe (e.g. Arnal et al., 2018; Harrigan et al., 2017). The basic principle behind the ESP is to create an ensemble of streamflows based on known initial conditions and forced by historic climate sequences. The historical record used in this study covers the period from 1980 to 2014 resulting in 34 members for each forecast. As in the hydrometeorological model chain, the ESP approach is set up in a leave one year out manner to ensure that the information of the year to be verified is not part of the forecasts itself. ESP predictions can be skillful especially in catchments where the predictability is mainly driven by the initial conditions, although the quality of the predictions depends on the seasons and hydroclimatic characteristics (Wood and Lettenmaier, 2008).

2.3 Hydrological Simulations

The hydrological simulations are performed with the Precipitation-Runoff-Evapotranspiration-Hydrotope model (PREVAH) (Gurtz et al., 1999; Viviroli et al., 2009a; Zappa et al., 2003). In this study, the distributed model version which requires gridded input data is used as described in Speich et al. (2015) and first applied in a hydrological study by Schattan et al. (2013). The model requires spatial information (land use, aspect and elevation) and gridded meteorological variables. Besides the statistically corrected temperature and precipitation predictions, relative humidity, sunshine duration, surface albedo and solar radiation are needed to run PREVAH. For the initialization of the model the required fields (i.e. soil moisture, groundwater storages, snow cover) are used from a reference simulation driven with the gridded meteorological observation dataset as described in subsection 3.4. The distributed hydrological model is run at a resolution of 200 m x 200 m, hence the same model internal procedures are used to further downscale the meteorological inputs for all different experiments. Namely, an interpolation based on inverse distance weighting (IDW) and different height and terrain specific correction are applied to the input variables (adiabatic lapse rate correction; aspect and slope corrections) as described in Zappa et al. (2003). A more extensive description of the model and a comparison ~~of~~ to the HRU based version of the model can be found in (Jörg-Hess et al., 2015a).

2.4 Verification

To verify the ~~streamflow~~ predictions various metrics are used to assess the forecast performance according to different characteristics or attributes of the forecasts respectively. The verification is performed for the two variables streamflow and snow water equivalent (SWE). For the selection of verifications metrics we follow the procedures presented in Brown et al. (2010) and we combine deterministic and probabilistic measures of skill to assess the forecast performance. More detailed reviews of ensemble forecast verification can be found in Jolliffe and Stephenson (2012) and Wilks (2006). Besides the Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970) widely used to assess the performance of hydrological models and the logarithmic version thereof (NSE log), which is more sensitive to low flows (Krause et al., 2005), we use the ensemble mean error (bias) and the mean absolute error (MAE). For both versions of the NSE and the bias the ensemble mean is used for the calculation. Although these metrics can describe the average characteristics of the ensemble forecasts it has been shown that it is crucial to consider the spread of the ensemble forecast as well to properly asses the forecast performance in particular at the subseasonal to seasonal time scale (Kumar et al., 2014). Therefore the continuous ranked probability skill score (CRPSS) with the extension proposed by Ferro (2014) to account for small ensemble sizes is used to characterize the performance of the full ensemble (Hersbach, 2000; Müller et al., 2005). In addition, we use the spread to error ratio (SprErr) as an indicator for the forecast reliability (Hopson, 2014). The *SprErr* is defined as the ratio between the variance of the forecast ensemble (forecast spread) and the mean squared error (MSE) of the ensemble forecast (forecast error). For reliable forecasts the spread and the error are equal, resulting in a SprErr of 1 whereas values below 1 indicate overconfidence (errors are larger compared to the spread) and values above 1 indicate overdispersion (the spread is larger compared to the error).

and Finally, rank histograms are used to visually examine the reliability of the forecasts (Anderson, 1996; Talagrand et al., 1997; Hamill & Colucci, 1997). In each of these scores the reforecasts are compared to observations (ME and MAE) and the climatology of the reference simulation is used as a reference climatology for NSE and CRPSS. In this study we use pseudo observations from a reference simulation to substitute real observations for the streamflow forecast verification. This is done to separate the effect on the performance of the hydrological predictions resulting from either different meteorological input forecasts or from the deficits in the hydrological model to simulate low flows. The setup of the reference simulation to generate the pseudo observations is described in the data subsection 3.4.

3. Data

3.1 Meteorological reforecast data

We obtained the subseasonal reforecasts from the ECMWF Integrated Forecasting System (IFS) version CY40r1. This version was operational from 19 November 2013 to 12 May 2015. This is a unique dataset because no system change occurred for nearly one and a half years (http://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/cycle-40r1/cycle-40r1 for the documentation of IFS CY40r1). Routinely the model is updated more frequently within one year and therefore changes in the system, e.g. changes to the horizontal and vertical resolution or changes in the parametrization of physical processes affect the skill of the forecast over the course of the year. The same data set is used as in the study presented by Monhart et al. (2018) and can be referred to for a more detailed description of the forecast system and for an extensive verification of these meteorological forecasts.

In this study the historical 5-member reforecasts (i.e. hindcasts) are used to drive the hydrological simulations. These meteorological reforecasts are run for all forecasts issued on Thursdays using ERA-interim analysis for the initialization (ECMWF, 2014) covering the period from April 1994 to March 2014 resulting in a total of 1040 individual reforecasts that are analyzed within this study. These reforecasts are essential for the post-processing of the meteorological forecasts (i.e. pre-processing from a hydrological point of view). The bias of the forecasting system can be estimated using the reforecasts and future forecasts can be corrected (or as in this analysis the reforecasts itself).

3.2 Meteorological observational data

We use gridded observation data sets at 2 km x 2 km resolution for daily temperature and precipitation to pre-process the meteorological reforecasts and for the verification thereof (Frei, 2014; Isotta et al., 2014; Meteoswiss, 2016; MeteoSwiss, 2016). These products are based on surface observations and are often used in climatological studies for Switzerland (e.g. Addor et al., 2016; Begert and Frei, 2018, Orth et al. (2015) for a hydrological model study with the same model version we use in this study). Nevertheless such gridded observational products exhibit limitations especially in complex terrain with high mountain peaks because of the interpolation errors and errors induced by natural variability (Frei, 2014; Addor & Fischer, 2015). Despite these limitations the analysis benefits by using the gridded version because of the scarcity of

meteorological stations available in the catchment areas. The additional meteorological variables (relative humidity, sunshine duration, surface albedo and solar radiation) needed to run the hydrological model PREVAH are directly used from the (raw) meteorological forecasts and are inherently downscaled with the PREVAH internal methodology. ~~Thus, no observations are needed for these variables.~~ Although these parameters could influence the hydrological relevant processes, e.g. evaporation rates from snow based on the relative humidity which could change after a downscaling, a thorough assessment of the effect of bias correcting and downscaling of these additional variables is out of scope of the current study. ~~In addition, these variables are of minor importance for the forecast time scales investigated in this study.~~

3.3 Catchment characteristics and hydrological data

The experiments analyzed in this study are performed for three alpine catchments with various hydro-climatic characteristics. This allows identifying the underlying processes that lead to skillful subseasonal streamflow predictions. The selection of the catchments is a compromise between the intended applications of our results within hydropower optimization in the Alpine region and to adequately meet the requirements of a scientific analysis. Hence, the Klöntal and the Verzasca catchment both are selected because of existing hydropower installation in these watersheds and the Thur catchment was chosen as a representative catchment with different hydroclimatic characteristics and because this catchment is often considered in hydrological research in Switzerland. An overview of the catchment characteristics can be found in Table 1.

The Verzasca catchment located in the Southern part of the Alps is snow dominated in the cold seasons and more precipitation dominated in warm seasons (Wöhling et al., 2006). The average height of the catchment is 1651 m a.s.l. with a maximum height of 2864 m a.s.l. and a minimum height at the gauging station of 490 m a.s.l.. Despite this height difference the catchment is of medium size only with a total of 185 km^2. The runoff in this catchment is low in winter due to accumulation of snow in elevated regions and highly dynamical in late spring to early autumn because the area is prone to heavy thunderstorm activity (Bogner et al., 2018; Liechti et al., 2013). The data used for calibration of the hydrological model and for the verification of the reference simulation is provided by the Federal Office for the Environment (FOEN) for the gauging station Lavertezzo.

The Klöntal catchment, located in the Northern part of the Alps, is mainly snow dominated and inherits a glaciated area of about 3 km^2 (< 5%). The maximum height is 2883m a.s.l. and the minimum height, corresponding to the height of lake Klöntal, is 847 m a.s.l. with a catchment area 83 km^2 categorized as a small catchment. The lake is used for hydropower generation. The data provided by the hydropower operator is used for calibration of the hydrological model and the verification of the streamflow reforecasts. Due to the absence of ~~a~~ gauging station~~s~~, this data is an estimate of the streamflow at the lake outlet and deduced from the water balance of historical lake levels. Hence, the quality of the data is lower compared to the other two catchments and therefore the interpretation especially under low flow condition should be treated with care.

The third catchment under investigation is the pre-alpine Thur catchment. This catchment is of medium size with 1696 km^2 and mainly precipitation dominated. The highest elevation is 2505 meter, the lowest elevation at the gauging station in Andelfingen provided by the FOEN is at 356 m, and the mean catchment height is 770 m a.s.l.. This catchment has frequently been analyzed in literature (Bogner et al., 2016; Fundel et al., 2013b; Jörg-Hess et al., 2015b).

3.4 Hydrological pseudo observations: the reference simulation

For all three catchments described above, the reference simulation is generated using an observed gridded meteorological data set for the meteorological variables (temperature, precipitation, relative humidity, sunshine duration, surface albedo and solar radiation) to run the hydrological model PREVAH (see section 2.3) in the same setting as used for different previous studies (Orth et al., 2015; Schattan et al., 2013; Speich et al., 2015). The gridding of all the meteorological variables makes use of elevation-based de-trending and inverse distance based interpolation as detailed in the baseline publication on PREVAH and its tools (Viviroli et al., 2009a) and also used for application of other hydrological models (e.g. Melsen et al., 2016). The reference simulation is a single time series starting in 1992 (after a spin up period of three years) and ending in 2015. The spin up period ensures the stability of the streamflow simulations and is of particular importance to fill low frequency storages for baseflow and snow (Viviroli et al., 2009b). In addition the spin-up times highly vary across catchments depending on the hydroclimatic characteristics and catchment size (Rahman et al., 2016; Seck et al., 2015). The climatology of the reference simulation is referred to as reference climatology and is used as reference for the computation of skill scores.

4. Results

4.1 Performance of the reference simulation

The reference simulation is analysed over the full 20-year period of the reforecasts. The verification of the reference simulation against observations, summarized in Table 2, generally shows high agreement of the simulation with the corresponding observations.

The Nash-Sutcliffe efficiency coefficient (NSE) for the entire analysis period (FullYear) exhibits values above 0.84 for all catchments. Best performance can be observed in spring and lowest performance in winter when low flows dominate. The logarithmic form of the NSE (NSE log) shows a similar picture with better performance in the Thur catchment. Some difficulties occur in DJF in the Verzasca catchment and in DJF and SON in the Klöntal catchment. In the Verzasca and the Klöntal catchment the MAE and the absolute bias are constant over the course of the year except in winter. But compared to the mean annual flows of the specific catchments largest relative biases are observed during DJF. The Thur catchment exhibits smallest relative biases except in JJA when low flows occur because of the absence of snow in this catchment. Various applications do not focus on low flows but rather on flow volumes, e.g. hydropower operations are more interested

in forecasts about the total upcoming flow volume to adapt and optimize their production. Therefore, we here focus on the standard Nash-Sutcliffe coefficient instead of the logarithmic form thereof.

We hereafter verify streamflow predictions against the reference simulation to focus on analyzing the effect of the different meteorological input forecasts. This allows to solely assess the effect of the pre-processing on the hydrological predictions by supressing the hydrological errors itself and is often done to evaluate ~~of~~ operational forecasting systems (Alfieri et al., 2014), to assess the contributions of uncertainty on the hydrological forecasts (Voisin et al., 2011), or for a comparison of the skill of different forecasting systems (Pappenberger et al., 2015). Such an evaluation against a reference simulation also minimizes the influence of the deficits of the hydrological model under low flow conditions on the verification analysis as found for the reference simulation described above. However, the effect of pre-processing on the performance if real observations are used for the verification reveals interesting aspects of the deficits of the hydrological simulations and will be discussed in section 5.

4.2 Skill of the meteorological input

The raw and bias corrected temperature and precipitation reforecasts used to drive the hydrological predictions are verified against the gridded observations in the Verzasca domain. In Figure 3, the CRPSS for the raw and the bias corrected temperature and precipitation reforecasts is shown. Raw temperature reforecasts mainly exhibit negative CRPSS values, indicating essentially no skill compared to climatology. After bias correction using the quantile mapping technique the reforecasts exhibit positive CRPSS up to 18-24 lead days in all seasons but spring (MAM). Raw precipitation reforecasts exhibit positive skill for lead days 5-11 in all seasons and negative CRPSS for longer lead times. After bias correction the skill is higher with positive CRPSS up to three weeks in ~~winter and spring~~MAM. In JJA the positive skill is only observed for 2 weeks lead time and in SON and DJF for the first week only. Similar results are obtained for the Klöntal and the Thur catchment (not shown).

4.3 Streamflow prediction performance

4.3.1 Prediction performance

The performance of the reforecasts is analysed for all available reforecast dates within the period 1994-2014. In Figure 4 the resulting scores (NSE, Bias, CRPSS and the spread-error-ratio) for all three catchments are presented indicating the skill of the prediction system against the reference simulation, i.e. the expected performance of the system for any date throughout the year.

For the Verzasca catchment, the skill in terms of the NSE from the ESP predictions suddenly drops after the initialization of the forecasts whereas the prediction system using meteorological reforecasts provides positive NSE up to 7 days. Pre-processing of precipitation generally even lowers the NSE indicating positive skill only for 5 lead days. Pre-processing of temperature does enhance the skill with positive NSE up to 13 days lead time. If both variables (precipitation and

temperature) are pre-processed the positive effect of the temperature pre-processing on NSE is diminished by the negative effect of precipitation pre-processing.

The negative biases of the ESP approach indicate an underestimation of the streamflows for all lead times in the Verzasca catchment. Raw forecasts show even stronger underestimation. After pre-processing either precipitation or temperature this underestimation is lower and reaches similar biases as the ESP predictions. If both parameters are pre-processed the biases are close to 0 or slightly positive indicating that the streamflows are neither under- nor overestimated.

The overall performance characterized by the CRPSS indicates positive skill for the ESP predictions up to 15 days lead time for the Verzasca, but the skill drops quickly after the initialization of the forecasts as in the NSE. Raw forecasts only show a positive CRPSS for the first 5 lead days. Pre-processing of precipitation increases the CRPSS at short lead times. Temperature pre-processing enhances the skill at early lead times and in addition elongates positive skill up to 15 days lead time. This is even more pronounced if both variables are pre-processed.

The spread error ratio of the ESP predictions is below 1 for all lead times indicating overconfidence. For the NWP-Hydro-chain the overconfidence is even higher for raw and temperature-only pre-processed (ppT) reforecasts. Pre-processed precipitation (ppP) reforecasts can partly correct the overconfidence of the streamflow reforecasts and if both variables are pre-processed (ppTP), the spread error ratio indicates reliable forecasts.

These skill signatures are similar in the other two catchments analyzed in this study (Figure 4), although not in an absolute sense. In the small, semi-glaciated Klöntal catchment the absolute skill generally is higher, and the skill of the predictions extends to longer lead times. In particular the raw and temperature-only pre-processed reforecasts (ppT) show positive skill in terms of the NSE throughout all lead times. In the Thur catchment the skill of the raw reforecasts outperforms the ESP predictions as well, but in contrast to both other catchments, the effect of pre-processing is negligibly small.

4.3.2 Seasonal variations in skill

The prediction skill of the different approaches does not only vary across catchments but as well across seasons. In Figure 5 the performance in the Verzasca catchment for the four seasons DJF, MAM, JJA and SON isare shown. The general characteristics are similar as observed for the entire year, i.e. the ESP based predictions exhibit a sudden drop in the NSE after initialization and most benefits (positive skill to longer lead times) are obtained if temperature-only is pre-processed (ppT). In DJF and MAM this skillful horizon is extended by ppT, from 5 (3) days to up to 16 (9) days in MAM (DJF) and the bias is reduced.

The overall performance (CRPSS) for the ESP predictions is better than the reference climatology for all lead times in MAM and clearly outperforms the raw reforecasts and precipitation-only pre-processed reforecasts (ppP) in MAM and DJF. For the temperature-only pre-processed reforecasts (ppT) and if both variables are pre-processed (ppTP), the predictions in these seasons (DJF and MAM) outperform the ESP forecasts for lead times up to 12 to 15 days and are equal for longer lead times. In JJA (and SON, not shown) the ESP predictions only exhibit a positive CRPSS up to 5 and 9 days, whereas the raw reforecasts (i.e. without any pre-processing) indicate positive skill up to lead times of 10 up to 15 days. However, in contrast

to DJF and MAM the influence of the pre-processing on the performance is negligible in JJA. Furthermore, in JJA the bias and the spread error ratio are only better if the pre-processing includes precipitation.

The seasonal variation in performance holds true as well for the other catchments, i.e. reforecasts initiated in winter and spring show highest benefits over the reference climatology. In contrast to the Verzasca catchment, the raw and precipitation only pre-processed reforecasts show higher streamflows than the reference climatology in MAM and the raw and temperature pre-processed reforecast show a positive NSE up to 30 lead time days (Figure 6, left).

Less seasonal variation is observed in the Thur catchment although the general signatures are evident as well. In MAM (Figure 6, right) all methods perform better than the reference simulation (positive CRPSS) over the full forecast range. Generally worst performance is found for the ESP. The effect of the pre-preprocessing is limited and can mainly be noticed if precipitation is pre-processed, resulting in smaller biases and a Spread to error ratio closer to 1, indicating a reduction in overconfidence.

4.3.3. Reliability of the ensembles

An additional important forecast characteristic is the reliability of the predictions, which cannot directly be deduced from the metrics shown above. Therefore, the rank histogram for the full period (Full year) and MAM reforecasts in the Verzasca catchment for all model configurations is shown in Figure 7 to assess the reliability of the streamflow forecasts. As an example we focus on the rank histograms of the full period (Full year) and the MAM reforecasts, because of its representativeness for the seasonality of performance. In MAM, both versions, raw and precipitation-only pre-processed reforecasts, show an underestimation of the flows (negative bias) indicating that most reforecast members tend to be lower than the corresponding observations. The strong negative bias is reduced if temperature-only pre-processed reforecasts (ppT) are used. But still a U-shape is evident in the histograms that indicates overconfidence and thus confirms the conclusions from the spread error ratio.

If both temperature and precipitation is pre-processed (ppTP) the resulting rank histograms become more uniform, in particular for longer lead times indicating a reduction of the overconfidence. But shorter lead times still exhibit some overconfidence. The rank histograms for the ESP predictions do provide more uniform rank histograms with a weak tendency of a negative bias. Although slight differences can be observed between different seasons, the main characteristics are similar for the full period.

The rank histograms for the Klöntal and Thur catchments (supplementary material) exhibit the same general behaviour regarding the reliability, but improvements by pre-processing are less pronounced for the Thur catchment and the rank histograms still indicate overconfidence even if both variables are pre-processed.

4.3.4 Snow water equivalent verification

To generate skillful streamflow predictions in mountainous catchments, the snow in the catchment is a crucial variable. Therefore, the snow water equivalent (SWE) in the hydrological model is analyzed according to different elevation regions.

For the verification we analyze the SWE at elevations above and below 1500 m a.s.l.. As in the verification of the streamflow reforecasts, the SWE is verified against the SWE of the reference simulation to replace the observations. ~~For an evaluation of the SWE against real observation the reader is referred to~~ A verification of modelled SWE against a consistent and homogenized climatology of gridded SWE based on station information is given by Jörg-Hess et al. (2014). They have shown that the modelled SWE exhibit errors that are in the same order as natural variability.

In MAM, raw and precipitation-only pre-processed (ppP) reforecasts highly overestimate the SWE in areas above 1500 m a.s.l. indicated by the MAE and the bias in Figure 8. The predictive skill in terms of the CRPSS is lost after 9 days lead time. In contrast, the reforecasts in DJF show stronger overestimation in areas below 1500 m a.s.l. and a total loss of predictive skill after 15 days lead time in this area. Lowest biases and highest skill (in terms of the CRPSS) is evident for reforecasts with pre-processed temperature and precipitation (ppTP), followed by temperature-only pre-processed (ppT) both outperforming the ESP predictions.

For all versions of the meteorological reforecasts (raw and pre-processed) the resulting SWE reforecasts tend to be overconfident, with least overconfidence if precipitation-only is pre-processed according to the spread error-ratio. The ESP predictions exhibit less overconfidence for both seasons and regions and exhibit similar levels in terms of MAE and the bias and slightly less overall skill (CRPSS) compared to ppT and ppTP reforecast versions.

The rank histograms confirm the conclusion drawn above. Raw and precipitation-only pre-processed reforecasts (ppP) exhibit largest positive biases throughout all lead times. In case of temperature-only (ppT) and temperature and precipitation pre-processed reforecasts (ppTP) the rank histograms indicate overconfidence in the beginning which is reduced for longer lead times.

The SWE verification in the Klöntal catchment shows a similar behavior with negative biases and largest MAE for the raw and precipitation-only pre-preprocessed (ppP) reforecasts. In contrast to the Verzasca catchment the CRPSS stays positive for all versions and for all lead times. A similar behavior is observed in the Thur catchment with positive skill for all lead times but smaller negative biases (the corresponding figures for the Klöntal and the Thur catchment can be found in the supplementary material).

5~~.~~ Discussion

For a proper evaluation of the effect of pre-processing on the hydrological streamflow predictions the following discussion considers the verification against the reference simulation. The meteorological input reforecasts highly benefit from the pre-processing procedure applied. The skill found for the pre-processed temperature and precipitation reforecast is comparable to the skill found in Monhart et al. (2018). In contrast to the present analysis, our earlier study used a station wise post-processing of the raw forecasts using the same setup as in the present study. Similarly, different studies emphasize the benefit of pre-processing precipitation (Crochemore et al., 2016) and temperature forecasts (Lucatero et al., 2017) at catchments at various spatial scales. The QM method used here is a popular pre-processing method for hydrometeorological

ensemble forecasts (e.g. Kang et al., 2010; Lucatero et al., 2018; Verkade et al., 2013) but does not come without limitations. In particular Zhao et al. (2017) point out the inability of QM to provide fully reliable ensembles for post-processing GCM precipitation. However, an extensive discussion of the reliability issue of the pre-processed meteorological input data used in this study can be found in Monhart et al. (2018). To summarize, it was found that the combination of the NWP model with QM indeed is able to provide reliable ensembles for lead times beyond 10 days but at shorter lead times the ensembles tend to be overconfident because the spread in the underlying NWP forecasts tends to be inappropriate what cannot be corrected using QM. An additional critical limitation of QM is the issue of variance inflation. Maraun (2013) emphasizes that the variance of the downscaled product strongly depends on the variance of the raw model grid box and QM does not introduce any small-scale variability. This is of particular importance for applications using local-scale information (such as distributed hydrological modelling) and if extremes are considered. In CH2018 (2018) these limitations of the QM method are highlighted for local climate change scenarios in Switzerland. In particular for convective precipitation events in summer the variance inflation issue can cause misinterpretation of data at the finer resolved scale. In the present study we are interested in the average streamflow throughout all season in the year for the upcoming 32 days and not in predicting extremes what reduces these implications, but still the spatial structure especially during convective situations in summer will likely be misrepresented and can influence the results. Different alternatives could be used depending on the specific application of the downscaled information, e.g. perfect prognosis approaches (Von Storch, 1999), the use of weather generators (e.g. Peleg et al., 2017) or in general stochastic methods (e.g. Volosciuk et al., 2017). However, such methods often require large computational resources. As the intention of this study is to pioneer the use of subseasonal hydrological predictions towards an operational use, we decided to use the QM technique despite its limitations. The results presented above and discussed in the following paragraphs legitimate our choice. Nevertheless, future studies should focus on the effect of variance inflation when QM is used to pre-process the predictions and alternative methods should be considered.

Our results show that subseasonal streamflow predictions in mountainous catchments can be skillful for the full 32 days lead time horizon in winter and spring. The traditional ESP approach clearly provides skillful predictions for all three catchments analyzed in this study, in terms of the overall skill (CRPSS) and the reliability. This is in agreement with the findings from Arnal et al. (2018) comparing the skill the of an ESP and a seasonal forecasting system across many regions in Europe. They found that the ESP approach can be outperformed mainly in the first month in terms of the CRPSS. Nevertheless, if scores evaluating the mean characteristics are considered (NSE and bias) we observe worse performance than suggested by the CRPSS. This indicates that the ESP predictions can capture the future evolution of the streamflow in a probabilistic sense. Furthermore, the substantial decrease in skill within the first days suggest that the ESP predictions are not able to capture the exact evolution but can rather be used to estimate the general behavior within the upcoming weeks. This is in agreement with the exponential decay in skill with increasing lead time found for ESP forecasts in UK catchments (Harrigan et al., 2017). If the NWP predictions are used to predict the streamflows, the skill can clearly be enhanced but in most cases only if the driving meteorological predictions are pre-processed. This indicates that the knowledge of the synoptic conditions plays an important role to enhance the skill of the streamflow predictions at early lead times and that biases in these driving

predictions need to be corrected prior to make these predictions useful. Furthermore, the improvements in performance of the streamflow predictions by pre-processing suggests that the gridded observational dataset provide a good baseline for this purpose, despite the difficulties involved in producing a gridded product based on a limited number of observational stations. In addition, the performance of the ESP predictions could potentially be enhanced if not all, but only individual years are taken into account. A certain guidance based on a selection of years with similar initial conditions could be taken into account. Crochemore et al. (2017) have shown that seasonal prediction based on ESP can benefit from condition the forecasts on climatology. However, an evaluation of such an approach is out of scope of the present study.

The effect of pre-processing is even more pronounced for the SWE verification. In the NWP chain the SWE is highly overestimated if temperature is not pre-processed (raw and ppP). Hence, the hydrological model inherent downscaling of temperature using an adiabatic laps rate leads to low skill in terms of CRPSS at longer lead times. Although, temperature lapse rate corrections have been found to be important for reproducing streamflows simulation based on regional climate model outputs in mountainous snow- and glacial dominated catchments (Butt and Bilal, 2011; Rahman et al., 2014) our study suggest that at least in a subseasonal forecasting context such corrections are not sufficient. Similarly, Tobin et al. (2011) have shown for flood forecasting framework constant lapse rate corrections even if seasonally-derived are unable to capture the dynamics of temperature changes during an event. At lower elevations this effect is even more crucial because the SWE is smaller and temperature biases accelerate melting processes in the model. If temperature pre-processing is included, these large errors can be avoided, and the skill of the SWE predictions is substantially increased. This effect underlines the importance of pre-processing the subseasonal forecast in snow dominated catchments. The importance of the SWE initial conditions for subseasonal forecasts has been shown by Jörg-Hess et al., (2015). They conclude that a better representation of snow melt process by improved states of the snow storage can greatly improve the predictions of streamflow volumes. The influence of initial conditions of SWE on seasonal streamflow predictions in the US is shown by Wood et al. (2016) in an idealized experiment. Furthermore, they stress that limited skill in seasonal meteorological predictions can be amplified in streamflow prediction skill. Our study suggests that an additional pre-processing of the meteorological forecasts is necessary to maintain the benefit of the initial conditions and confirms the findings of amplified skill in the streamflow predictions if the forecasts exhibit some skill. Otherwise, low skill in performance of the (raw) forecasts and the loss in skill to predict the SWE over longer lead times directly translates into low skill of the streamflow forecasts.

The comparison of the performance in the three different catchments analyzed gives further insight in the predictability of streamflows in alpine catchments and strengthens the picture drawn above which shows the importance of the complex interactions between precipitation, temperature and SWE at the subseasonal forecast time scale. In the Thur catchment, the largest catchment in the analysis, which is mainly precipitation dominated, the difference in skill between the ESP predictions and the NWP based prediction is smallest. Similarly, the differences between raw and pre-processed experiments are marginal. Only in spring, if precipitation is pre-processed (ppP and ppTP) a small improvement in the CRPSS can be observed at early lead times coinciding with the lead times where pre-processed precipitations still exhibit skill. Hence the

initial conditions of the model and skillful precipitation predictions at early lead times determine to a large degree the skill of the reforecasts and the negligible snow-covered area within the catchment does marginally affect the performance of the runoff predictions. In the Thur catchment two additional subcatchments in the upper Thur catchment (at the runoff stations Halden and Murg) have been verified to identify the influence of hilliness and catchment size on the forecast performance. Although the size of the Murg (212 km^2) and the Halden catchment (~~1750~~ 1085 km^2) varies significantly, the hilliness in both is comparable, while the lower part of the Thur catchment (station Andelfingen, 1696 km^2) is a typical lowland region. For all three stations the skill is very similar (supplementary material Figure S4) and pre-processing does not vary either. This suggest that neither hilliness nor catchment size does significantly influence the performance of the forecast. In the snow dominated and partially glaciated Klöntal catchment, the smallest catchment in this analysis, the high skill in terms of the NSE indicates a good performance of the ensemble mean but in terms of the overall skill (CRPSS) the NWP prediction are only skillful if temperature pre-processing is considered. This superiority in the mean is most likely the effect of melting processes. On the other hand, the Verzasca catchment is snow dominated in elevated regions, but rain dominated at lower elevations due to its large gradient of elevations. Therefore, snow-melt processes tend to occur more often and are more diverse due to the higher elevation gradient within the catchment. The NSE does indicate skill up to 13 days lead time for the temperature-only pre-processed forecast (ppT) but precipitation pre-processing even lowers the skill despite an increase in skill of the corresponding precipitation inputs. This counterintuitive behavior of lowering the skill (in terms of NSE) in the streamflow prediction despite the use of improved precipitation inputs underlines that for a profound assessment of the skill of ensemble forecasts, verification metrics focusing on mean flows can be insufficient and misleading.

The combination of pre-processed subseasonal meteorological prediction with hydrological simulations can outperform a traditional ESP approach in small to medium-sized alpine catchment. Especially in snow dominated and (semi-) glaciated catchments such a prediction chain brings large benefits in the forecast performance. But temperature (and precipitation) from the NWP model needs to be pre-processed prior to be used in hydrological models to achieve better performance than an ESP approach. In precipitation dominated catchment the pre-processing does only show a marginal improvement in skill, but the NWP chain clearly outperforms ESP predictions. Hence, such systems can be of interest for application when accurate and reliable runoff predictions are desired, especially in snow-dominated catchments. Furthermore, Frei et al., (2018) have shown that a general decrease of snow fall is expected in future climate change scenarios, while at higher elevation the signal shows a slight increase in heavy snowfall events, due to a shift of climatological cold areas into a temperature interval which favor higher snowfall intensities in combination with a general increase in winter precipitation. Especially regarding future scenarios which expect an increase in hydro power production due to melt water in the period from October to April (Weingartner et al., 2013) such systems might become a valuable tool for optimizing hydropower production in mountainous areas. Future work should include statistical post-processing techniques (of the hydrological output) to correct the errors and biases of the hydrological simulation and to account for additional uncertainty induced by the hydrological model in the ensemble.

The discussion above focused on the effect of pre-processing hydrometeorological predictions and therefore only the verification against the reference simulation was considered and hydrological model errors are thus excluded from the analysis. However, to estimate the real world performance the hydrological model errors need to be taken into account. To do so we verified the streamflows of the reference simulation of the PREVAH model against observations. The evaluation presented in section 4.1 revealed the good performance of the hydrological model with NSE above 0.8 in most seasons. Largest difficulties are observed in seasons with low flow conditions (in DJF in the Verzasca, and in SON and DJF in the Klöntal catchment). This is particularly evident in the logarithmic version of the NSE, in which flood peaks are flattened to better asses the performance under low-flow conditions, the difficulties of the model remain. These deficits of the hydrological model need to be considered in the verification of the predictions and are the reason for verifying the reforecasts against the pseudo observations from the reference simulation. Otherwise, if the predictions are verified against real streamflow observations the hydrological model deficiencies dominate the skill characteristics of the predictions and possibly impede the identification of the effect of pre-processing. An example of such a deficiency is the uncertainty resulting from the hydrological modelling that result in stronger overconfidence especially at short lead times when the hydrological model uncertainty may be the dominant source of uncertainty as discussed for example in Bennett et al. (2014). To illustrate this for the prediction used in this study, the verification was repeated with the real runoff observations. The skill in most seasons (MAM, JJA, SON) exhibits the same behavior with high skill at early lead times and decreasing skill at longer lead times and generally higher overconfidence at short lead times is observed what confirms the findings by Bennett et al. (2014). But during low flow condition (in DJF in the Verzasca catchment and in DJF and SON in the Klöntal catchment), strong negative peaks in skill (in terms of the CRPSS) are present at short lead times, with increasing skill at longer lead times (supplementary material). The corresponding input data revealed that this behavior can be associated with predicted snow melt events that coincide with observed “no melt” events. During such events, all members of the experiments including pre-processed temperatures (ppT ad ppTP) overestimate the observed runoff peak. The comparison of the meteorological input data from the reforecast with data from an observational station shows an overestimation in both temperature and precipitation. The temperature observations for this event were clearly below freezing, whereas the raw forecast data is close to 0° C, and the pre-processed temperature is positive. Hence, the melt-affected area is too large leading to an overestimation of the runoff contribution. In addition, overestimated precipitation will further contribute to runoff and less deposition of solid precipitation will occur. Similarly, the reference simulation does overestimate the runoff peak because of an overestimation of the temperature in the gridded data used to run the hydrological model. It is known that gridded observations do inherit errors due to interpolation (Freudiger et al., 2016) and do have difficulties in resolving small-scale features, such as for example cold pools in alpine valleys (Frei, 2014). This can partly explain this behavior. Another potential explanation is an insufficient formulation of the discrimination between rain and snow in the hydrological model. In the version of PREVAH used for this study, the formulation follows a threshold based method in combination with a linear range as described in Zappa et al. (2003). This linear transition range is set between the threshold values -1.5°C and +1.5°C ~~set to +/- 1.5°C~~ as determined by the hydrological calibration. The threshold itself and the corresponding linear range

highly depends on the hydroclimatic characteristics and thus can strongly vary in space (Liu et al., 2018). It has been shown that more sophisticated approaches using logistic regression for characterizing this range can provide better results (Frei, 2016). In principle such an approach could be included into the hydrological model, but such an implementation is out of scope of this analysis. Alternatively, it has been shown that such errors of the hydrological model can be corrected by additionally post-processing the hydrological output using neuronal networks or logistic regressions (e.g. Bogner et al., 2016; Sharma et al., 2018).

6 Conclusion and outlook

Recent advances in subseasonal meteorological ensemble models makes it feasible to develop hydrometeorological prediction systems driven by such NWP forecasts. We developed an end-to-end hydrometeorological prediction system driven with reforecasts from the subseasonal prediction system from the ECMWF. A pre-processing procedure based on QM is used to bias correct and downscale the meteorological predictions prior to the hydrological model. The performance of the resulting streamflow forecasts is assessed for three small to medium size alpine catchments using various verification metrics to assess different attributes of the reforecasts. Our study demonstrates the potential of ensemble streamflow predictions in small mountainous catchments. Moreover, the benefits of combining NWP predictions and hydrological models has been shown. The analysis indicates the need for pre-processing of the driving meteorological prediction especially in small snow-dominated catchments in alpine regions.

Decent skill of traditional ESP predictions compared to climatological reference can extend up to the entire 32 days. The NWP approach outperforms the E~~P~~SP predictions in all catchments and most seasons in particular at short lead times up to about day 5. In snow dominated catchments, an additional pre-processing step of both~~,~~ temperature and precipitation~~,~~ is crucial to further enhance the skill and the reliability of the forecasts. While pre-processing precipitation-only is not sufficient to enhance the forecast performance, it is crucial in the combination with temperature preprocessing to improve the forecast reliability. Again, it is noted here that the verification is done against the reference simulation as replacement of real observations. Hence, the performance cannot directly be interpreted as the prediction performance in an absolute sense, because in our approach the deficits of the hydrological model are not fully taken into account. However, the relative benefits of using NWP output as forcing for the hydrological simulations and the improvements after pre-processing is expected to hold true as well with real observations.

The benefits of the NWP approach and the pre-processing step is most pronounced in winter and spring when snow melt processes dominate. This demonstrates the importance of snow for the predictability of streamflows at the subseaonal timescale. Hence the deficits in the hydrological model with respect to snow related processes (in particular the distinction between solid and liquid precipitation) should receive further attention to enhance the forecast performance. Alternatively, post-processing techniques applied to the streamflow forecasts can be applied to correct such hydrological model deficits. This would allow assessing the skill of the forecasts with respect to real observations and can potentially further increase the

performance of the forecasts. Hence, if post-processing techniques are able to account for the deficits of the reference simulation, the combination of both pre- and post-processing could provide skillful lead streamflow predictions in snow and glacial dominated catchment in mountainous terrain at the subseasonal forecast horizon.

Furthermore, technical improvements of the NWP models related to the ensemble size, frequency of issuing reforecasts and improvements in the representation of physical processes, can be expected to have a positive effect on the resulting streamflow performance. In our setup we use a rather simple statistical bias correction technique to pre-process the hydrometetroloigcal prediction. More sophisticated pre-processing techniques could be applied to analyze their capability to enhance the streamflow performance. Since ensemble hydrometeorological predictions are of interest for specific applications the forecasts should further be analyzed according to their economic value. For example to optimize the revenues of existing hydropower plants in alpine regions or for early better preparedness of hydrological droughts.

Data availability:

ECMWF forecast data are accessible through the MARS archive: http:/ /apps.ecmwf.int/archive-catalogue/, observation records used in this study are available from the following website: https://gate.meteoswiss.ch/idaweb/more.do (SwissMetNet). Streamflow series and catchment boundaries are provided by the Swiss Federal Office for the Environment (FOEN).

References

Addor, N. and Fischer, E. M.: The influence of natural variability and interpolation errors on bias characterization in RCM simulations, J. Geophys. Res. Atmos., 120(19), 10,180-10,195, doi:10.1002/2014JD022824, 2015.

Addor, N., Rohrer, M., Furrer, R. and Seibert, J.: Propagation of biases in climate models from the synoptic to the regional scale: Implications for bias adjustment, J. Geophys. Res. Atmos., 121(5), 2075–2089, doi:10.1002/2015JD024040, 2016.

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D. and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, J. Hydrol., 517, 913–922, doi:10.1016/j.jhydrol.2014.06.035, 2014.

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Clim., 9(7), 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, 1996.

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B. and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrol. Earth Syst. Sci., 22(4), 2057–2072, doi:10.5194/hess-22-2057-2018, 2018.

Bauer, P., Thorpe, A. and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, doi:10.1038/nature14956, 2015.

Beerli, R., Wernli, H. and Grams, C. M.: Does the lower stratosphere provide predictability for month-ahead wind electricity generation in Europe?, Q. J. R. Meteorol. Soc., 143(709), 3025–3036, doi:10.1002/qj.3158, 2017.

Begert, M. and Frei, C.: Long-term area-mean temperature series for Switzerland—Combining homogenized station data and high resolution grid data, Int. J. Climatol., 38(6), 2792–2807, doi:10.1002/joc.5460, 2018.

Bennett, J. C., Robertson, D. E., Lal, D., Wang, Q. J., Enever, D., Hapuarachchi, P. and Tuteja, N. K.: A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days, J. Hydrol., 519, 2832–2846, doi:10.1016/j.jhydrol.2014.08.010, 2014.

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.

Bogner, K., Liechti, K., Bernhard, L., Monhart, S. and Zappa, M.: Skill of Hydrological Extended Range Forecasts for Water Resources Management in Switzerland, Water Resour. Manag., 32(3), 969–984, doi:10.1007/s11269-017-1849-5, 2018.

Bohn, T. J., Sonessa, M. Y. and Lettenmaier, D. P.: Seasonal Hydrologic Forecasting: Do Multimodel Ensemble Averages Always Yield Improvements in Forecast Skill?, J. Hydrometeorol., 11(6), 1358–1372, doi:10.1175/2010JHM1267.1, 2010.

Brown, J. D., Demargne, J., Seo, D. J. and Liu, Y.: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environ. Model. Softw., 25(7), 854–872, doi:10.1016/j.envsoft.2010.01.009, 2010.

Butt, M. J. and Bilal, M.: Application of snowmelt runoff model for water resource management, Hydrol. Process., 25(24), 3735–3747, doi:10.1002/hyp.8099, 2011.

CH2018: CH2018 - Climate Scenarios for Switzerland, Technical Report, National Center for Climate Services, Zurich. [online] Available from: https://www.nccs.admin.ch/nccs/de/home/klimawandel-und-auswirkungen/schweizer-klimaszenarien/technical-report.html, Chapter 5.7., p98-101, 2018.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol., 375(3–4), 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

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(9), 3601–3618, doi:10.5194/hess-20-3601-2016, 2016.

Crochemore, L., Ramos, M.-H., Pappenberger, F. and Perrin, C.: Seasonal streamflow forecasting by conditioning climatology with precipitation indices, Hydrol. Earth Syst. Sci., 21(3), 1573–1591, doi:10.5194/hess-21-1573-2017, 2017.

Davolio, S., Diomede, T., Marsigli, C., Miglietta, M. M., Montani, A. and Morgillo, A.: Comparing different meteorological ensemble approaches: hydrological predictions for a flood episode in Northern Italy, Adv. Sci. Res., 8, 33–37, doi:10.5194/asr-8-33-2012, 2012.

Day, G. N.: Extended Streamflow Forecasting Using NWSRFS, J. Water Resour. Plan. Manag., 111(2), 157–170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157), 1985.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D. J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J. and Zhu, Y.: The science of NOAA's operational hydrologic ensemble forecast service, Bull. Am. Meteorol.

Soc., 95(1), 79–98, doi:10.1175/BAMS-D-12-00081.1, 2014.

Díez, E., Primo, C., García-Moya, J. A., Gutiérrez, J. M. and Orfila, B.: Statistical and dynamical downscaling of precipitation over Spain from DEMETER seasonal forecasts, Tellus, Ser. A Dyn. Meteorol. Oceanogr., 57(3), 409–423, doi:10.1111/j.1600-0870.2005.00130.x, 2005.

Van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J. and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resour. Res., 49(5), 2729–2746, doi:10.1002/wrcr.20251, 2013.

Ferro, C. A. T.: Fair scores for ensemble forecasts, Q. J. R. Meteorol. Soc., 140(683), 1917–1923, doi:10.1002/qj.2270, 2014.

Frei, C.: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances, Int. J. Climatol., 34(5), 1585–1605, doi:10.1002/joc.3786, 2014.

Frei, P.: Snowfall in the Alps: Evaluation and Projections based on the EURO-CORDEX Regional Climate Models, Sci. Rep. Meteoswiss, (101), 122, 2016.

Frei, P., Kotlarski, S., Liniger, M. A. and Schär, C.: Future snowfall in the Alps: Projections based on the EURO-CORDEX regional climate models, Cryosphere, 12(1), 1–24, doi:10.5194/tc-12-1-2018, 2018.

Freudiger, D., Frielingsdorf, B., Stahl, K., Steinbirch, A., Weiler, M., Griessinger, N. and Seibert, J.: The Potential of meteorological gridded datasets for hydrological modeling in alpine basins, Hydrol. und Wasserbewirtschaftung, 60(6), 353–367, doi:10.5675/HyWa_2016,6_1, 2016.

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, 2013a.

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, 2013b.

García-Morales, M. B. and Dubus, L.: Forecasting precipitation for hydroelectric power management: how to exploit GCM's seasonal ensemble forecasts, Int. J. Climatol., 27, 1691–1705, doi:10.1002/joc.1608, 2007.

Gurtz, J., Baltensweiler, A. and Lang, H.: Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins, , 13(March), 2751–2768, 1999.

Hamill, T. M. and Colucci, S. J.: Verification of Eta – RSM Short-Range Ensemble Forecasts, Mon. Weather Rev., 125, 1312–1328, 1997.

Hao, Z., Singh, V. P. and Xia, Y.: Seasonal Drought Prediction: Advances, Challenges, and Future Prospects, Rev. Geophys., 56(1), 108–141, doi:10.1002/2016RG000549, 2018.

Harrigan, S., Prudhomme, C., Parry, S., Smith, K. and Tanguy, M.: Benchmarking Ensemble Streamflow Prediction skill in the UK, Hydrol. Earth Syst. Sci. Discuss., (July), 1–28, doi:10.5194/hess-2017-449, 2017.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, 2000.

Hock, R.: Temperature index melt modelling in mountain areas, J. Hydrol., 282(1–4), 104–115, doi:10.1016/S0022-1694(03)00257-9, 2003.

Hopson, T. M.: Assessing the Ensemble Spread – Error Relationship, Mon. Weather Rev., (Wilks 2011), 1125–1142, doi:10.1175/MWR-D-12-00111.1, 2014.

IFS documentation ECMWF: Part V : Ensemble Prediction System IFS DOCUMENTATION – Cy40r1 Operational implementation 22 November 2013 PART V : ENSEMBLE PREDICTION SYSTEM, , (November 2013), 1–25, 2014.

Inman, R. H., Pedro, H. T. C. and Coimbra, C. F. M.: Solar forecasting methods for renewable energy integration, Prog. Energy Combust. Sci., 39(6), 535–576, doi:10.1016/j.pecs.2013.06.002, 2013.

Isotta, F. A., Frei, C., Weilguni, V., Perčec Tadić, M., Lassègues, P., Rudolf, B., Pavan, V., Cacciamani, C., Antolini, G., Ratto, S. M., Munari, M., Micheletti, S., Bonati, V., Lussana, C., Ronchi, C., Panettieri, E., Marigo, G. and Vertačnik, G.: The climate of daily precipitation in the Alps: Development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data, Int. J. Climatol., 34(5), 1657–1675, doi:10.1002/joc.3794, 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. Siences, 8, 281–291, 2008.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification, edited by I. T. Jolliffe and D. B. Stephenson, John Wiley & Sons, Ltd, Chichester, UK., 2011.

Jörg-Hess, S., Fundel, F., Jonas, T. and Zappa, M.: Homogenisation of a gridded snow water equivalent climatology for Alpine terrain: Methodology and applications, Cryosph., 8(2), 471–485, doi:10.5194/tc-8-471-2014, 2014.

Jörg-Hess, S., Griessinger, N. and Zappa, M.: Probabilistic Forecasts of Snow Water Equivalent and Runoff in Mountainous Areas, J. Hydrometeorol., 16, 2169–2186, doi:10.1175/JHM-D-14-0193.1, 2015a.

Jörg-Hess, S., Kempf, S. B., Fundel, F. and Zappa, M.: The benefit of climatological and calibrated reforecast data for simulating hydrological droughts in Switzerland, , 22, 444–458, doi:10.1002/met.1474, 2015b.

Kang, T., Kim, Y. and Hong, I.: Comparison of pre- and post-processors for ensemble streamflow prediction, Atmos. Sci. Lett., 159(June), 153–159, doi:10.1002/asl.276, 2010.

Krause, P., Boyle, D. P. and Bäse, F.: Advances in Geosciences Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5(89), 89–97, doi:10.5194/adgeo-5-89-2005, 2005.

Kumar, S., Dirmeyer, P. A. and Kinter, J. L.: Usefulness of ensemble forecasts from NCEP climate forecast system in sub-seasonal to intra-annual forecasting, Geophys. Res. Lett., 41(10), 3586–3593, doi:10.1002/2014GL059586, 2014.

Van Lanen, H. A. J., Laaha, G., Kingston, D. G., Gauster, T., Ionita, M., Vidal, J., Vlnas, R., Tallaksen, L. M., Stahl, K., Hannaford, J., Delus, C., Fendekova, M., Mediero, L., Prudhomme, C., Rets, E., Romanowicz, R. J., Gailliez, S., Wong, W. K., Adler, M., Blauhut, V., Caillouet, L., Chelcea, S., Frovola, N., Gudmundsson, L., Hanel, M., Haslinger, K., Kireeva, M., Osuch, M., Sauquet, E., Stagge, J. H. and Van Loon, A. F.: Hydrology needed to manage droughts : the 2015 European case, Hydrol. Process., 3104(April), 3097–3104, doi:10.1002/hyp.10838, 2016.

Lavers, D., Luo, L. and Wood, E. F.: A multiple model assessment of seasonal climate forecast skill for applications,

Geophys. Res. Lett., 36(23), L23711, doi:10.1029/2009GL041365, 2009.

Liechti, K., Panziera, L., Germann, U. and Zappa, M.: The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps, Hydrol. Earth Syst. Sci., 17(10), 3853–3869, doi:10.5194/hess-17-3853-2013, 2013.

Liu, Y., Ren, G., Sun, X. and Li, X.: A new method to separate precipitation phases, Hydrol. Earth Syst. Sci. Discuss., (July), 1–38, doi:10.5194/hess-2018-307, 2018.

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., (July), 1–28, doi:10.5194/hess-2017-366, 2017.

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J. and Jensen, K. H.: Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: The effect of preprocessing and post-processing on skill and statistical consistency, Hydrol. Earth Syst. Sci., 22(7), 3601–3617, doi:10.5194/hess-22-3601-2018, 2018.

Manzanas, R., Gutiérrez, J. M., Fernández, J., van Meijgaard, E., Calmanti, S., Magariño, M. E., Cofiño, A. S. and Herrera, S.: Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications, Clim. Serv., 9, 44–56, doi:10.1016/j.cliser.2017.06.004, 2018.

Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, J. Clim., 26(6), 2137–2143, doi:10.1175/JCLI-D-12-00821.1, 2013.

Marsigli, C., Montani, A. and Paccagnella, T.: A spiatial verification method applied to the evaluation of high-resolution ensemble forecasts, Meteorol. Appl., 15, 125–143, doi:10.1002/met.65, 2008.

Meißner, D., Klein, B. and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe, Hydrol. Earth Syst. Sci., 21(12), 6401–6423, doi:10.5194/hess-21-6401-2017, 2017.

Melsen, L., Teuling, A., Torfs, P., Zappa, M., Mizukami, N., Clark, M. and Uijlenhoet, R.: Representation of spatial and temporal variability in large-domain hydrological models: Case study for a mesoscale pre-Alpine basin, Hydrol. Earth Syst. Sci., 20(6), 2207–2226, doi:10.5194/hess-20-2207-2016, 2016.

Meteoswiss: Documentation of MeteoSwiss Grid-Data Products Daily Mean , Minimum and Maximum Temperature: TabsD , TminD , TmaxD, Available Meteoschweiz Kundend. [online] Available from: https://www.meteoswiss.admin.ch/home/services-and-publications/produkte.subpage.html/en/data/products/2014/raeumliche-daten-temperatur.html, 2016.

MeteoSwiss: Documentation of MeteoSwiss grid-data products: Daily precipitation (final analysis): RhiresD, Available Meteoschweiz Kundend. [online] Available from: http://www.meteoswiss.admin.ch/home/services-and-publications/produkte.subpage.html/en/data/products/2014/raeumliche-daten-niederschlag.html, 2016.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C. and Liniger, M. A.: Skill of Subseasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations, J. Geophys. Res. Atmos., 123(15), 7999–8016, doi:10.1029/2017JD027923, 2018.

Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T. and Buizza, R.: Nonlinear Processes in Geophysics Performance of

the ARPA-SMR limited-area ensemble prediction system : two flood cases, Nonlinear Process. Geophys., 8(6), 387–399, doi:10.5194/npg-8-387-2001, 2001.

Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T., Tibaldi, S. and Buizza, R.: The Soverato flood in Southern Italy: performance of global and limited-area ensemble forecasts, Nonlinear Process. Geophys., 10, 261–274, 2003.

Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T.: Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: Main achievements and open challenges, Tellus, Ser. A Dyn. Meteorol. Oceanogr., 63(3), 605–624, doi:10.1111/j.1600-0870.2010.00499.x, 2011.

Müller, W. A., Appenzeller, C., Doblas-Reyes, F. J. and Liniger, M. A.: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes, J. Clim., 18(10), 1513–1523, doi:10.1175/JCLI3361.1, 2005.

Nash, E. and Sutcliffe, V.: River flow forecasting through conceptual models Part1 - A discussion of preinciples, J. Hydrol., 10, 282–290, 1970.

Ohmura, A.: Physical Basis for the Temperature-Based Melt-Index Method, J. Appl. Meteorol., 40(4), 753–761, doi:10.1175/1520-0450(2001)040<0753:PBFTTB>2.0.CO;2, 2001.

Orth, R. and Seneviratne, S. I.: Predictability of soil moisture and streamflow on subseasonal timescales: A case study, J. Geophys. Res. Atmos., 118(19), 10963–10979, doi:10.1002/jgrd.50846, 2013a.

Orth, R. and Seneviratne, S. I.: Propagation of soil moisture memory to streamflow and evapotranspiration in Europe, Hydrol. Earth Syst. Sci., 17(10), 3895–3911, doi:10.5194/hess-17-3895-2013, 2013b.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J. and Zappa, M.: Does model performance improve with complexity? A case study with three hydrological models, J. Hydrol., 523, 147–159, doi:10.1016/j.jhydrol.2015.01.044, 2015.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A. and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, J. Hydrol., 522, 697–713, doi:10.1016/j.jhydrol.2015.01.024, 2015.

Peleg, N., Fatichi, S., Paschalis, A., Molnar, P. and Burlando, P.: An advanced stochastic weather generator for simulating 2-D high-resolution climate variables, J. Adv. Model. Earth Syst., 9(3), 1595–1627, doi:10.1002/2016MS000854, 2017.

Rahman, K., Etienne, C., Gago-Silva, A., Maringanti, C., Beniston, M. and Lehmann, A.: Streamflow response to regional climate model output in the mountainous watershed: a case study from the Swiss Alps, Environ. Earth Sci., 72(11), 4357–4369, doi:10.1007/s12665-014-3336-0, 2014.

Rahman, M. M., Lu, M. and Kyi, K. H.: Seasonality of hydrological model spin-up time: a case study using the Xinanjiang model, Hydrol. Earth Syst. Sci. Discuss., (July), 1–22, doi:10.5194/hess-2016-316, 2016.

Robertson, A. W., Kumar, A., Peña, M. and Vitart, F.: Improving and Promoting Subseasonal to Seasonal Prediction, Bull. Am. Meteorol. Soc., ES53(March), 49–53, doi:10.1175/BAMS-D-14-00139.1, 2015.

Schaake, J. C., Hamill, T. M., Buizza, R. and Clark, M.: HEPEX: The hydrological ensemble prediction experiment, Bull. Am. Meteorol. Soc., 88(10), 1541–1547, doi:10.1175/BAMS-88-10-1541, 2007.

Schär, C., Vasilina, L., Pertziger, F. and Dirren, S.: Seasonal Runoff Forecasting Using Precipitation from Meteorological

Data Assimilation Systems, J. Hydrometeorol., 5(5), 959–973, doi:10.1175/1525-7541(2004)005<0959:SRFUPF>2.0.CO;2, 2004.

Schattan, P., Zappa, M., Lischke, H., Bernhard, L., Thurig, E. and Diekkruger, B.: An approach for transient consideration of forest change in hydrological impact studies, IAHS-AISH Proc. Reports, 359(November 2015), 311–319, 2013.

Seck, A., Welty, C. and Maxwell, R. M.: Spin-up behavior and effects of initial conditions for an integrated hydrologic model, Water Resour. Res., 51(4), 2188–2210, doi:10.1002/2014WR016371, 2015.

Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza, P. and Mejia, A.: Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system, Hydrol. Earth Syst. Sci., 22(3), 1831–1849, doi:10.5194/hess-22-1831-2018, 2018.

Sinha, T. and Sankarasubramanian, A.: Role of climate forecasts and initial conditions in developing streamflow and soil moisture forecasts in a rainfall-runoff regime, Hydrol. Earth Syst. Sci., 17(2), 721–733, doi:10.5194/hess-17-721-2013, 2013.

Speich, M. J. R., Bernhard, L., Teuling, A. J. and Zappa, M.: Application of bivariate mapping for hydrological classification and analysis of temporal change and scale effects in Switzerland, J. Hydrol., 523, 804–821, doi:10.1016/j.jhydrol.2015.01.086, 2015.

Von Storch, H.: On the use of "inflation" in statistical downscaling, J. Clim., 12(12), 3505–3506, doi:10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2, 1999.

Talagrand, O., Vautard, R. and Strauss, B.: Evaluation of probabilistic prediction systems, in Proceedings of a Workshop held at ECMWF on Predictability, 20-22 October 1997, Reading, United Kingdom: European Center for Medium-Range Weather Forecasts, pp. 1–25., 1997.

Thirel, G., Martin, E., Mahfouf, J. F., Massart, S., Ricci, S., Regimbeau, F. and Habets, F.: A past discharge assimilation system for ensemble streamflow forecasts over France - Part 2: Impact on the ensemble streamflow forecasts, Hydrol. Earth Syst. Sci., 14(8), 1639–1653, doi:10.5194/hess-14-1639-2010, 2010.

Tobin, C., Nicotina, L., Parlange, M. B., Berne, A. and Rinaldo, A.: Improved interpolation of meteorological forcings for hydrologic applications in a Swiss Alpine region, J. Hydrol., 401(1–2), 77–89, doi:10.1016/j.jhydrol.2011.02.010, 2011.

Verbunt, M., Walser, A., Grutz, J., Montani, A. and Schär, C.: Probabilistic Flood Forecasting with a Limited-Area Ensemble Prediction System: Selected Case Studies, J. Hydrometeorol., 8, 897–909, doi:10.1175/JHM594.1, 2007.

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, J. Hydrol., 501, 73–91, doi:10.1016/j.jhydrol.2013.07.039, 2013.

Vitart, F. and Robertson, A. W.: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events, npj Clim. Atmos. Sci., 1(1), 3, doi:10.1038/s41612-018-0013-0, 2018.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Dï¿½quï¿½, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H. S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I.,

Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D. J., Xiao, H., Zaripov, R. and Zhang, L.: The subseasonal to seasonal (S2S) prediction project database, Bull. Am. Meteorol. Soc., 98(1), 163–173, doi:10.1175/BAMS-D-16-0017.1, 2017.

Viviroli, D., Zappa, M., Gurtz, J. and Weingartner, R.: An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools, Environ. Model. Softw., 24(10), 1209–1222, doi:10.1016/j.envsoft.2009.04.001, 2009a.

Viviroli, D., Zappa, M., Schwanbeck, J., Gurtz, J. and Weingartner, R.: Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland - Part I: Modelling framework and calibration results, J. Hydrol., 377(1–2), 191–207, doi:10.1016/j.jhydrol.2009.08.023, 2009b.

Voisin, N., Pappenberger, F., Lettenmaier, D. P., Buizza, R. and Schaake, J. C.: Application of a Medium-Range Global Hydrologic Probabilistic Forecast Scheme to the Ohio River Basin, Weather Forecast., 26(4), 425–446, doi:10.1175/WAF-D-10-05032.1, 2011.

Volosciuk, C., Maraun, D., Vrac, M. and Widmann, M.: A combined statistical bias correction and stochastic downscaling method for precipitation, Hydrol. Earth Syst. Sci., 21(3), 1693–1719, doi:10.5194/hess-21-1693-2017, 2017.

Weingartner, R., Schädler, B. and Hänggi, P.: Auswirkungen der klimaänderung auf die schweizerische wasserkraftnutzung, Geogr. Helv., 68(4), 239–248, doi:10.5194/gh-68-239-2013, 2013.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B. and Zebiak, S. E.: Potential applications of subseasonal-to-seasonal (S2S) predictions, Meteorol. Appl., 24(3), 315–325, doi:10.1002/met.1654, 2017.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 3rd ed., edited by D. S. Wilks, International Geophysics Series, Vol. 100, Academic Press Inc., London., 2011.

Wöhling, T., Lennartz, F. and Zappa, M.: Technical Note: Updating procedure for flood forecasting with conceptual HBV-type models, Hydrol. Earth Syst. Sci., 10(6), 783–788, doi:10.5194/hess-10-783-2006, 2006.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35(14), 1–5, doi:10.1029/2008GL034648, 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.

Zappa, M., Pos, F., Strassera, U., Warmerdam, P. and Gurtzl, J.: Seasonal Water Balance of an Alpine Catchment as Evaluated by Different Methods for Spatially Distributed Snowmelt Modelling, Nord. Hydrol., 34(3), 179–202, 2003.

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E. and Ramos, M. H.: How suitable is

quantile mapping for postprocessing GCM precipitation forecasts?, J. Clim., 30(9), 3185–3196, doi:10.1175/JCLI-D-16-0652.1, 2017.

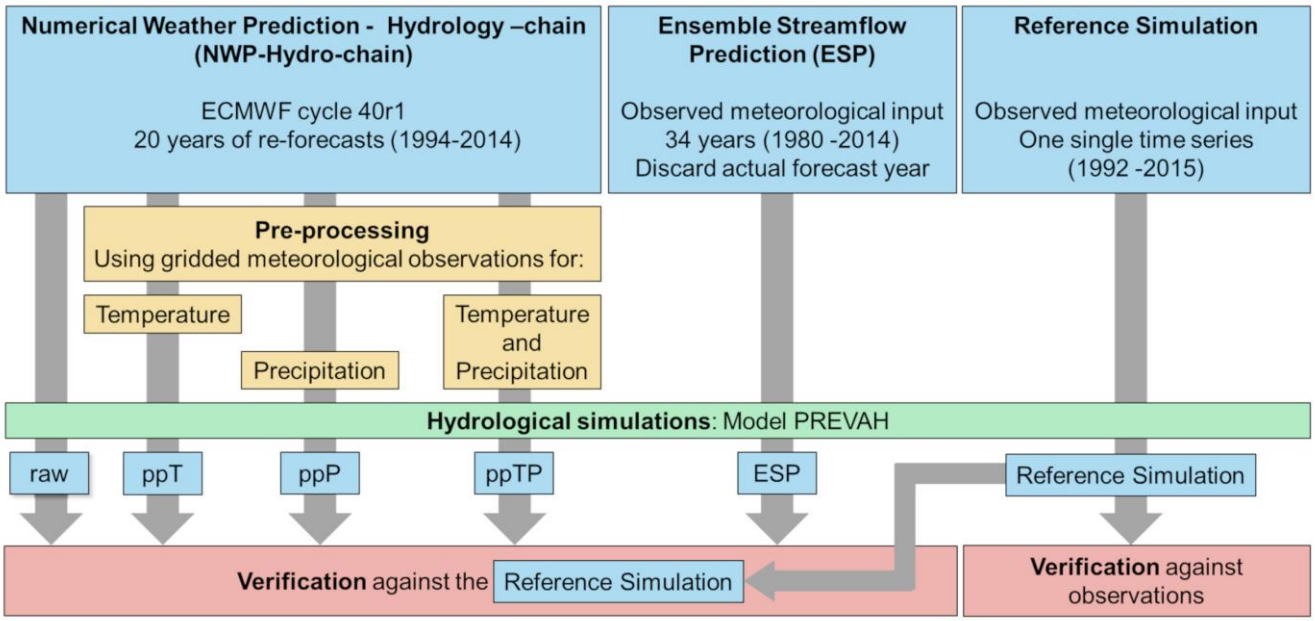

Figure 1: Conceptual design of the NWP-Hydro-chain with and without pre-processing (left) and the ESP chain (right).

5

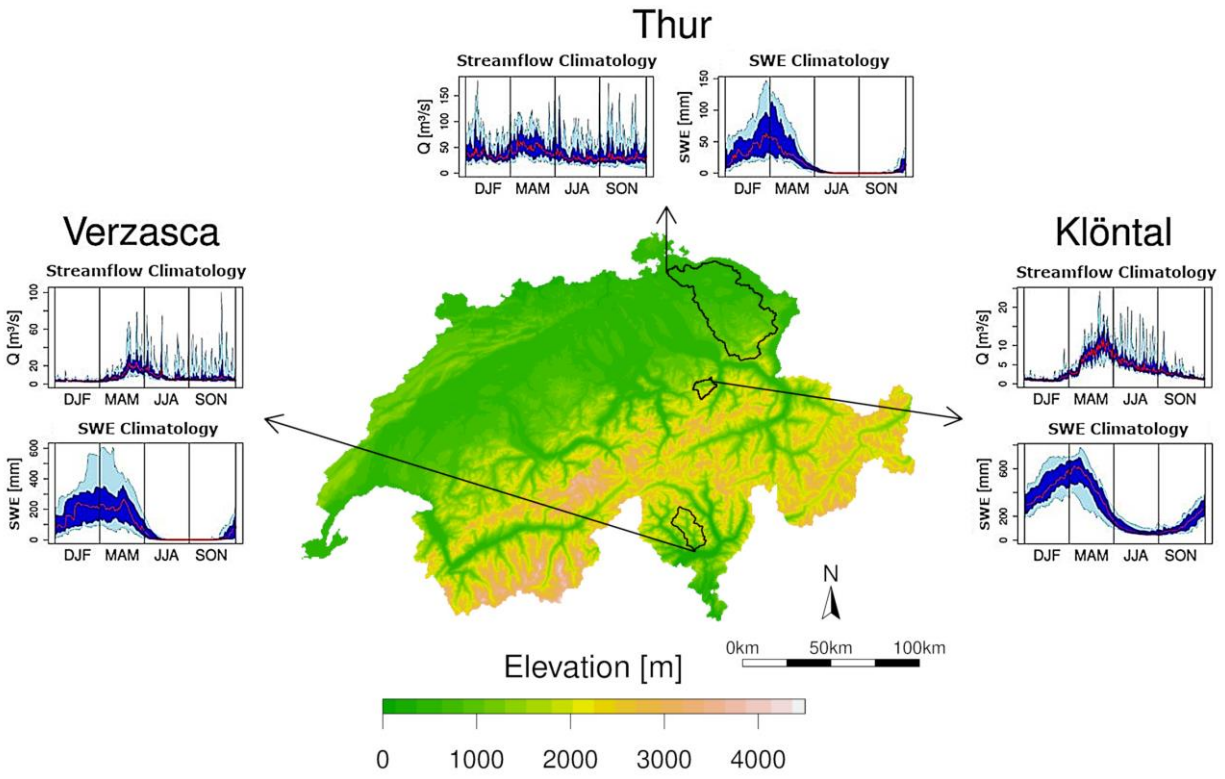

Figure 2: Location and extent of the three selected catchments and their corresponding streamflow and SWE climatology. The Verzasca catchment in the south, Klöntal catchment in the center and the Thur catchment in the north east.

		Verzasca	Klöntal	Thur
Catchment area [km2]		185	83	1696
altitude range [m]	maximum elevation	2864	2883	2505
	average elevation	1651	*1704~~NA~~*	770
	minimum elevation	490	847	356
dominant hydroclimatic regime		snow	snow and glacial	precipitation

Table 1: Overview of the catchment characteristics for the three catchments analysed.

		Verzasca		Klöntal		Thur	
NSE	**FullYear**	**0.85**		**0.84**		**0.84**	
	DJF	0.38		0.39		0.83	
	MAM	0.88		0.82		0.85	
	JJA	0.82		0.75		0.79	
	SON	0.84		0.75		0.86	
NSE log	**FullYear**	**0.87**		**0.68**		**0.87**	
	DJF	0.44		-0.14		0.82	
	MAM	0.90		0.84		0.87	
	JJA	0.86		0.78		0.85	
	SON	0.89		0.43		0.88	
MAE [m^3/s] ([%])	**FullYear**	**2.84**	**(25.9)**	**1.19**	**(26.6)**	**10.54**	**(22.5)**
	DJF	1.35	(36.1)	0.73	(50.7)	10.89	(25.2)
	MAM	3.16	(18.7)	1.20	(19.1)	11.53	(16.2)
	JJA	3.40	(33.9)	1.63	(26.7)	11.30	(28.7)
	SON	3.45	(23.1)	1.21	(39.6)	8.50	(23.9)
Bias [m3/s] ([%])	**FullYear**	**0.28**	**(2.5)**	**0.26**	**(5.9)**	**-0.42**	**(-0.9)**
	DJF	1.07	(28.7)	0.39	(27.2)	2.49	(5.8)
	MAM	-0.06	(-0.4)	-0.01	(-0.2)	-1.47	(-2.1)
	JJA	0.23	(2.3)	0.04	(0.6)	-4.41	(-11.2)
	SON	-0.11	(-0.7)	0.66	(5.9)	1.74	(4.9)

Table 2: Verification of the reference simulation with corresponding observations for the Verzasca catchment, the Klöntal catchment and the Thur catchment. In each catchment the Nash-Sutcliffe coefficient (NSE), the Nash-Sutcliffe coefficient using logarithmic input values (NSE log), the bias and the mean absolute error are shown. The verification is done for the full simulation period (FullYear: 1994-2014) and the individual seasons within the full period (DJF, MAM, JJA, SON). A perfect simulation would have NSE=1, and positive values of NSE indicated better skill than the reference climatology

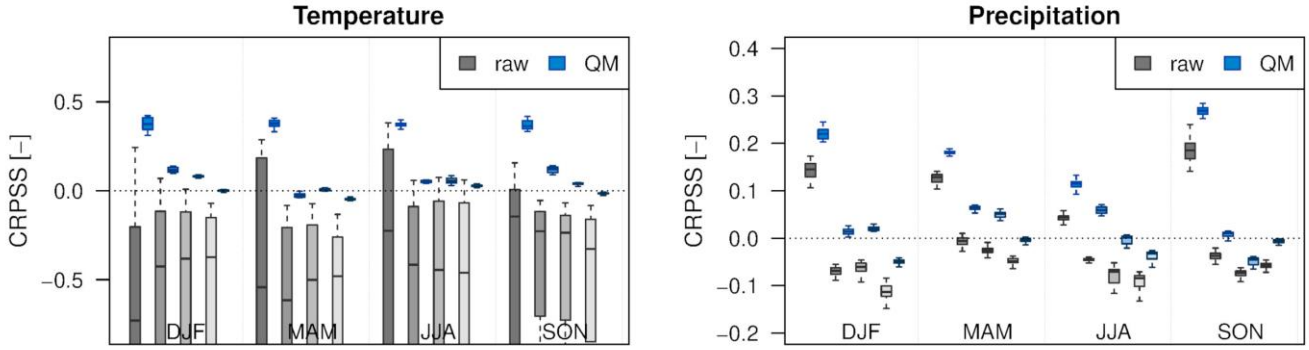

Figure 3: Overview of the CRPSS for weekly mean temperatures (left) and weekly precipitation sums (right) in 1994-2014 reforecasts grouped by season (DJF, MAM, JJA and SON) for all grid points within the Verzasca catchment. The shading of the boxes denotes lead time, whereas week 1 corresponds to day 5-11, week 2 to days 12-18 and so on. Grey shading for raw forecast and blue shading for pre-processed reforecasts. An individual box shows the distribution of the CRPSS for all grid points within the catchment averaged over all 13 reforecast initialization dates within one season. The boxes depict the interquartile range, the mean is indicated by the horizontal line and the whiskers span the length of 1.5 x standard deviation of the data. A perfect forecast has CRPSS=1, and positive values indicate better skill than the reference climatology.

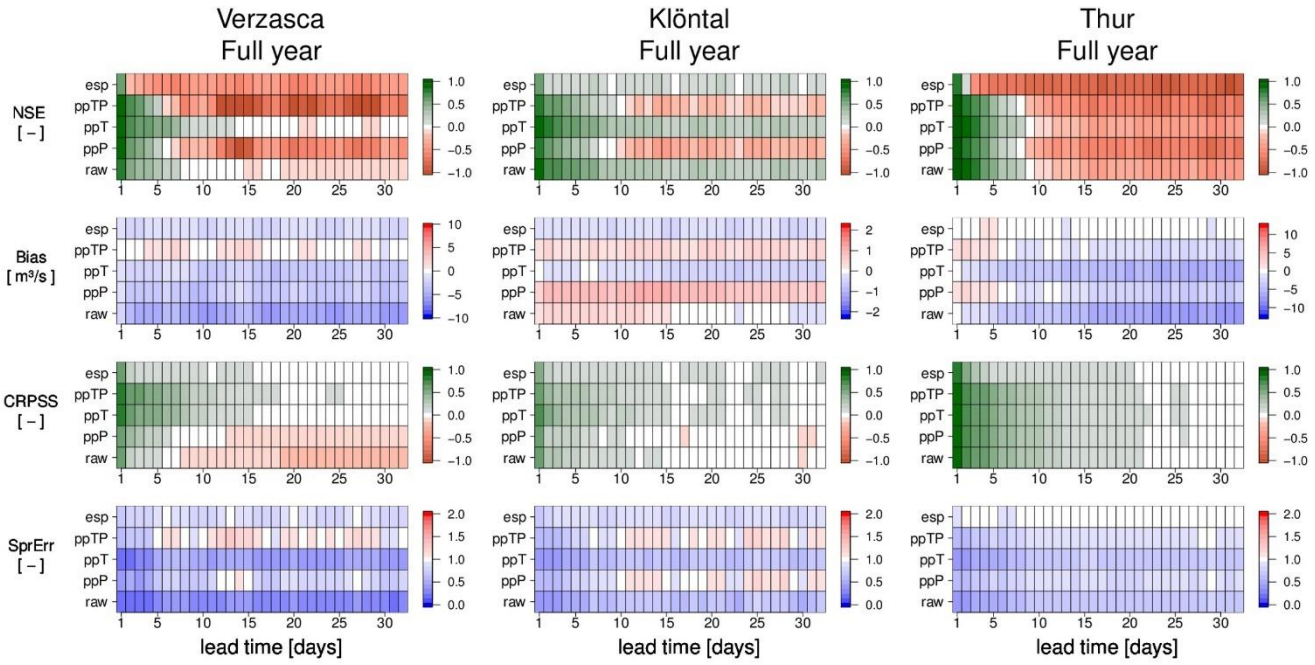

Figure 4: Verification of the streamflow forecasts for the Verzasca (left), the Klöntal (middle) and the Thur catchment (right) considering all forecasts within the reforecast period. The Nash-Sutcliffe coefficient (NSE) in the upper most panel, the mean bias (Bias) in second, the CRPSS in the third, and the Spread-Error relationship (SprErr) in the lowest panel. For each score the 5 different setups are shown. The first row corresponds to the ESP approach (esp), the second row to the reforecasts using both pre-processed temperature and precipitation (ppTP), the third row to the pre-processed temperature-only (ppT), fourth row to the pre-processed precipitation-only (ppP) and the fifth row to the reforecasts using the raw meteorological input reforecasts (raw). The NSE and the CRPSS span from -infinty to 1 with a perfect score being 1; a bias of zero indicates no forecast error with negative values indicating underestimation and positive values indicating overestimation of the flow; reliable forecasts exhibit a SprErr of 1 and lower values indicate overconfidence and greater values indicate overdispersion.

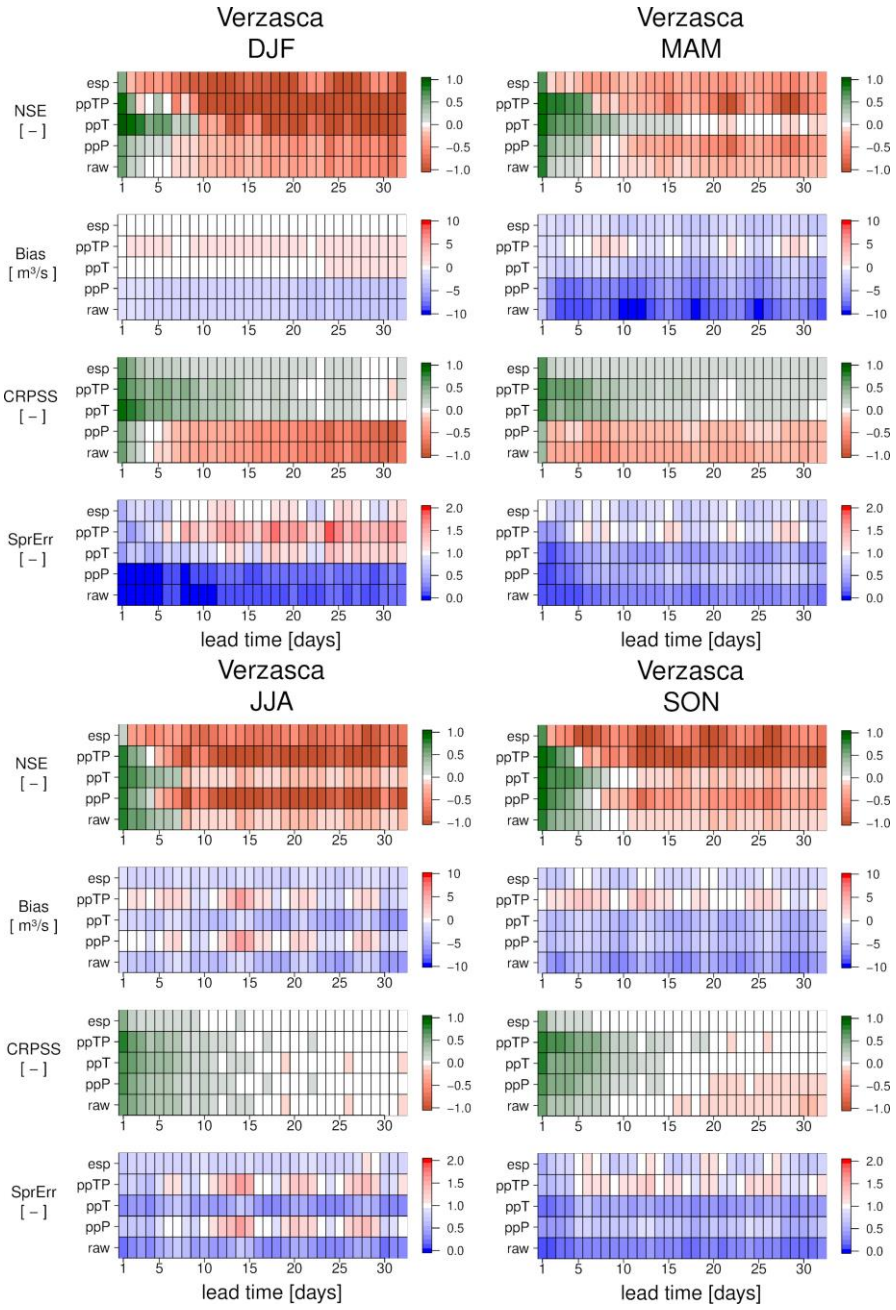

Figure 5: Same as Figure 4 but seasonally aggregated (DJF, MAM, JJA and SON) for the Verzasca catchment.

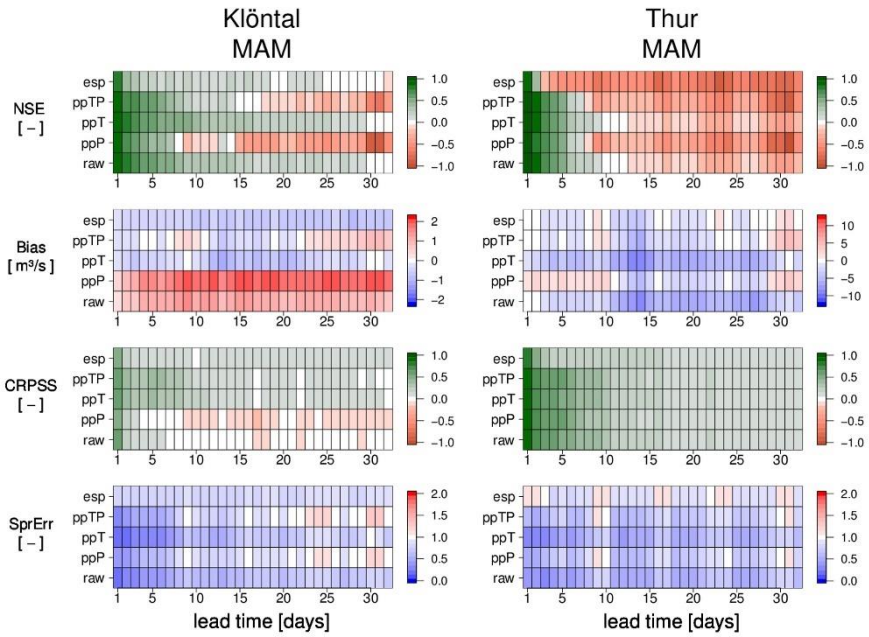

Figure 6: Same as Figure 4 but for the Klöntal and the Thur catchment in MAM.

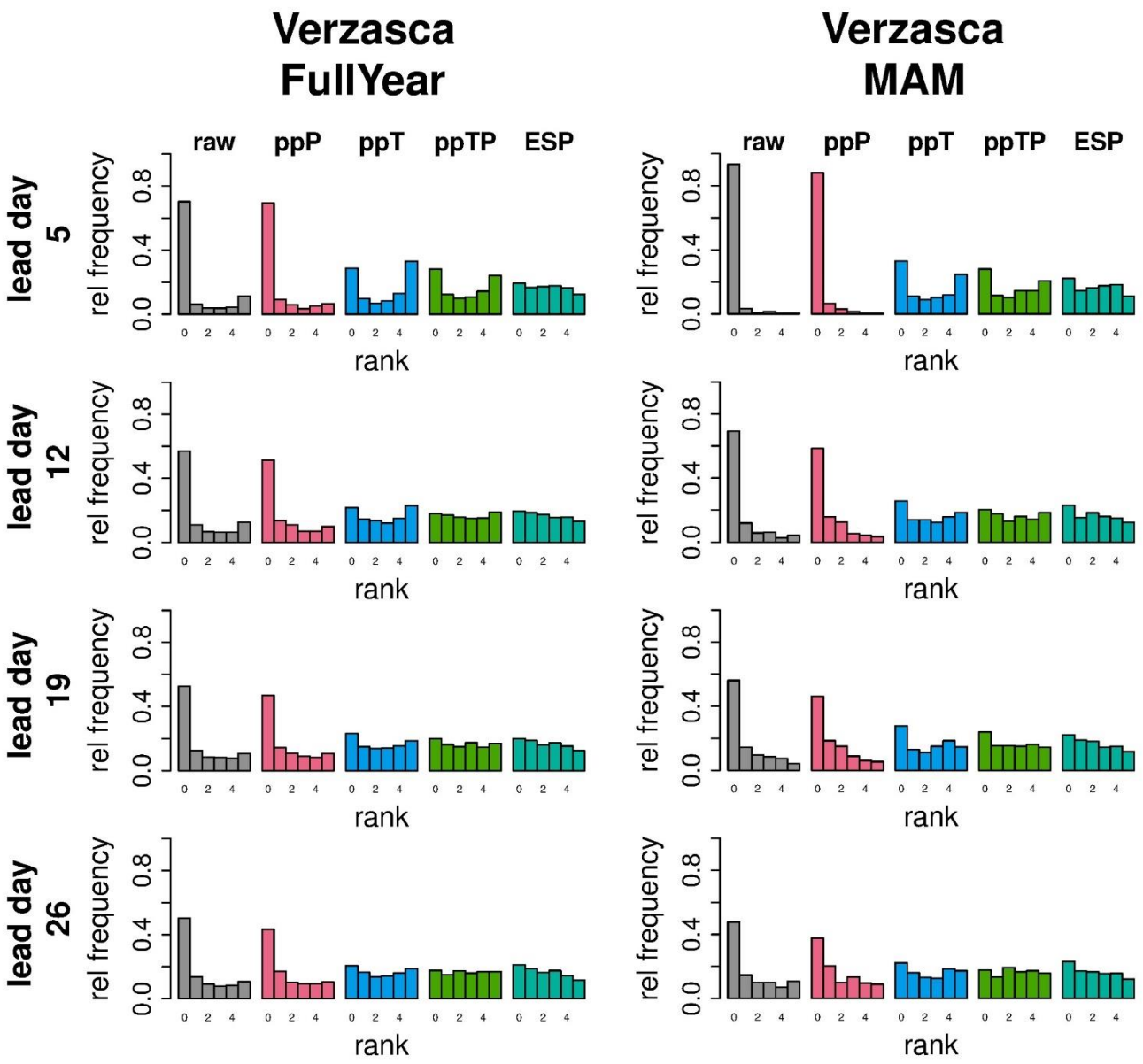

Figure 7: Rank histograms for all four configurations (raw, ppP, ppT, ppTP) and the ESP predictions for the full analysis period (Full year) and MAM in the Verzasca catchment. The basic principle in the rank histograms the assumption that the ensemble members determine bins in which the corresponding observation can be ranked. For a reliable forecast, the observations are equally distributed across all different bins resulting in a uniform shape of the rank histograms. If for example the frequency in the lowest (rank 0) and the highest bin (rank 6) is much higher, the observation tends to be more frequently either higher than all ensemble members or lower than all ensemble members but less often in between the ensemble members. This specific U-shape indicates that the forecast spread is too narrow and thus the forecasts generally overconfident. In contrast, if the observations tend to be more often in between the ensemble members (e.g. rank 2 and 3) the rank histogram exhibit a convex shape and thus the forecast spread is too large indicating overdispersive forecasts.

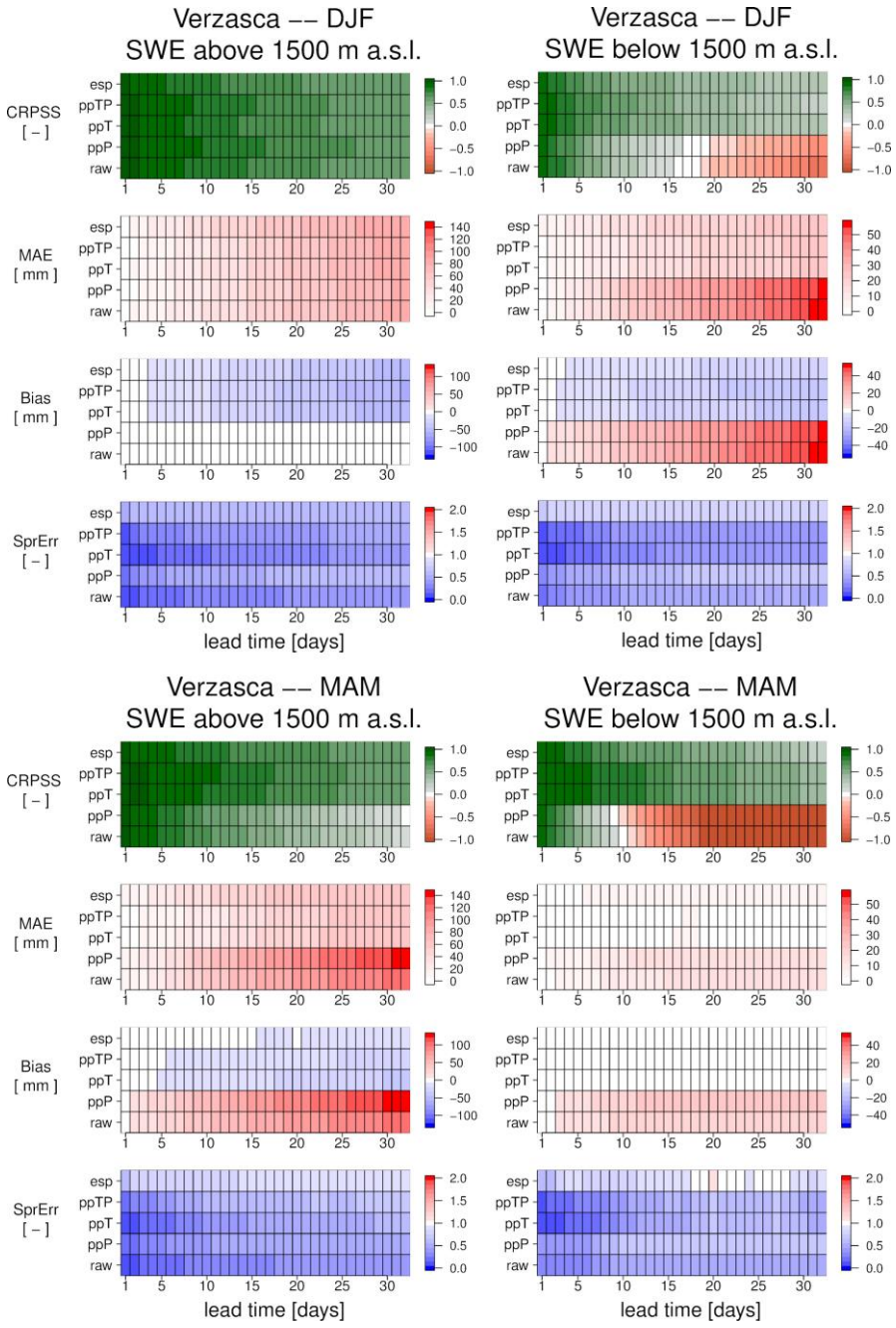

Figure 8: Same as in Figure 5 but for SWE in DJF and MAM in the Verzasca catchment.