Response to reviewer 1

Reviewer's comments are in blue, our comments are in black.

General remarks

This paper describes results from a hindcast study of a multimodel approach for ensemble streamflow forecasting. While the method has been described in an earlier paper, this paper focuses on the performance evaluation of two variations: MEads and MEhds. The skill metrics are well explained and results show a clear improvement of skill compared to the historical ensemble (HE, or classical ESP) method.

The text is well-written and clear. Below are a few minor comments and questions.

Questions and comments

Page 3, line 30: It surprises me that no weighting scheme was applied. I would expect you have knowledge about the relative performance of the different model chains from experience of from earlier studies (Olson et al, 2016). Is this something you intend to investigate in the future?

Yes, we understand your surprise here. The wording used (page 4, lines 1-2) implies we did not test weighting schemes which is not the case. We did in fact test two types of weighting, a simple arithmetic weighting system similar to that used by Olsson et al. (2016) and a linear regression based approach. In both cases the sharpness of the multi-model ensemble was improved significantly but there was no improvement in the multi-model's general performance. The pooling approach was still able to hold a small advantage over the weighted versions.

The following changes were made to correct and clarify what we did:

page 4, lines 1-2 was changed to read,

"The simple weighting scheme used by Olsson et al. (2016) was tested but, other than improving the ensemble sharpness, did not offer an improvement over the pooling approach."

The following paragraph has been added to the conclusion as well:

"How the individual model ensembles are combined to give the multi-model output needs to be revisited. When we applied the asymmetric weighting scheme proposed by Olsson et al. (2016) we did not find that it improved the multi-model performance in general across all stations and forecasts and so did not use it. However, we do believe that more work should be done to find a more appropriate weighting scheme than simple pooling. Perhaps by better understanding how the

performance of the different modelling chains are affected by the initial conditions and lead-time it will shed more light on how to best approach this issue. Further development and testing along these lines are planned for the future."

Page 4, line 29: Do you use all index values from this period, or the average, or else?

We use the mean or average index values for the period. We have reworded the sentences (page 4, 29) to clarify this,

"The teleconnection indices they identified are the Arctic oscillation (AO) and the Scandinavian pattern (SCA) and the periods of persistence for these indices, expressed as the index mean for the identified period, are the seven and eight months leading up to the spring flood respectively."

Page 10, line 21 and 22: "starting from 1961" and "period 1961-2015" Is this a typo? On page 7, line 2, a different period was mentioned (1981-2015). If this is not a typo, why can the years 1961-1980 not be used for the performance evaluation?

This is a typo. The reason that the entire data series cannot be used is that the hindcasts of the driving data used in the multi-model are only available from 1981.

We have changed it to read, "…the data used in our work are for the period 1981-2015 due to some of the other datasets used in this work only being available from 1981".

Grammar and spelling

Use punctuation when using adjuncts, for example on page 2:

- To achieve this, operators …
- In practice, there are …

The text has been reviewed and punctuation added when using adjuncts as suggested.

Page 10, line 6: "subbasins sub-basins"

The second hyphenated word has been removed.

Page 10, line 17: "(hereafter SFV)" (this abbreviation has been introduced before).

Changed to read, "We focus on forecasts of the accumulated streamflow volume during this period or SFV."

Page 12, line 20: "… to perform the better than…"

The word 'the' has been removed so that it now reads, "… to perform better than…"

Page 13, line 16: "an mean"

The ‘an’ has been changed to an ‘a’ to read, “a mean”.

Page 14, line 17: “with regards to”

The phrase has been replaced with the ‘at’ so the sentence reads, “Out of the three terciles the prototype shows the least skill over HE at discriminating between NN events and non-NN events.”

Page 15, line 17: “however the prototype”

The word ‘the’ has been added as suggested.

Table 3: LT, MT and UT are not introduced. Do you mean BN, NN, AN?

Yes, this is a typo and we did mean BN, NN, AN. The change has been made.

Response to reviewer 2

Reviewer's comments are in blue, our comments are in black.

Brief Overview

The paper presents a very interesting study related to the implementation of a prototype for seasonal forecasting in Swedish rivers based on hydrological modelling and seasonal meteorological forecasts. The prototype is compared to a traditional operational EPS approach and to climatology. Results show benefits in the use of the prototype. The paper is well written, methods are adequately described, and assessments seems suitable to the objectives. I have only a few major and minor comments about the manuscript.

Major Comments

P2, l20-25: Please observe that it historical observations are referred two times. And only in the second one it is presented as the ESP approach. The explanation here could be better.

We have reworded page 2, lines 19 and line24, to make this clearer. They now read as follows:

"…and then force it with either historical observations (called ensemble streamflow prediction or ESP; e.g. Day, 1985)…"

"Another dynamical approach is the well-established ESP method (Day, 1985)."

Evaluation section: I understand that one of the limitations of the work is that authors were not able to evaluate properly the ensemble, since most of the used metrics are related to transforming the ensemble into the ensemble mean, and then evaluating it as a deterministic forecast. Authors did not even experiment testing some other metrics?

With only 35 data points per station, one data point per year, we felt that it was not enough data on which to perform a robust probabilistic evaluation on. We experimented with the metric CRPS but were ultimately uncomfortable presenting those results due to their uncertainty arising from the limited data used in the analysis. We should also point out that the inter quartile range skill score (IQRSS) and uncertainty sensitivity skill score (USS) used in this work are basic ensemble evaluation metrics and, although not a full probabilistic evaluation, do give some insight into the performance of the forecast ensembles.

P5, l10: It is relevant to better explain what is the data used in the bias correction. Also, I think this procedure has great impact in results, but it is not adequately described. My suggestion is to explore more this point.

We have expanded our description of the bias adjustment and have replaced page 5, line 10 with the following:

"A change to previous work has these daily P and T data bias adjusted first before being used to force HBV. The bias adjustment method used is a version of the distribution based scaling approach (DBS; Yang et al., 2010) which has been adapted for use on seasonal forecast data. DBS is a quantile mapping bias adjustment method where meteorological variables are fitted to appropriate parametric distributions (e.g. Berg et al., 2015; Yang et al., 2010). For precipitation, two discrete gamma distributions are used to adjust the daily seasonal forecast values, one for low-intensity precipitation events (≤ 95th percentile) and another for extreme events (> 95th percentile). For temperature, a Gaussian distribution is used to adjust the daily seasonal forecast values.

Observed (Sect. 2.6 Study area and local data) and seasonal forecast (Sect. 2.7 Driving Data) time-series of P and T spanning the relevant forecast timeframe (e.g. Jan-Jul for forecasts initialised in January) and for the reference period 1981-2010 are used to derive the adjustment factors to transform the seasonal forecast data to match the observed frequency distributions. First the precipitation data is adjusted then the temperature data. The latter is done separately for dry and wet days in an attempt to preserve the dependence between P and T (e.g. Olsson et al. 2010; Yang et al, 2010). Adjustment factors are calculated for each calendar month as the distributions can have different shapes depending on the physical characteristics of the precipitation processes that are dominant. It should be emphasized that the adjustment parameters were estimated using much of the same data to which they were applied. Ideally the parameters would be estimated using data that does not overlap the data which is being adjusted. However, this was not possible in the scope of this work."

Conclusions:

Authors commented that the prototype was put into operation as a beta product at SMHI in January 2017. This gives openness for another discussion: in an operational perspective, are the benefits verified

for the prototype enough to justify the implementation? I understand that yes, but also the prototype is more dependent on data and require more processing power and time to run, right?

Yes, we and the power companies think that they do. It must be emphasised that every percent improvement in the forecast error can potentially be converted into large financial revenues for the power companies and energy traders. So an average improvement in forecast error, over all subbasins and initialisation dates, by 6% (individual results can be as high as 31%, see figure 4) can be viewed as a significant improvement. Care was taken while developing the prototype to minimise the added computational power and data requirements. Additionally, these forecasts are made only once a month so the additional computational time, ca. 1 extra hour, is not a significant factor.

Page 16, line 6 was rewritten to emphasise that the implementation of the prototype as a beta product was done together with the power companies. It now reads:

"These results have been met with great interest from the hydropower industry and the prototype was put into operation, in cooperation with the power companies, as a beta product at SMHI in January 2017."

Minor Comments

P1, l16: "considered" is doubled in the text

The first instance has been deleted so that it now reads, "Both the considered multi-model methods considered showed skill over the reference forecasts…"

P2, l34: Please explain better what is "limited success". Only one case is cited

The paper cited is a review of the different experiments performed at SMHI to improve the forecast error of the SFV. They found that a despite these efforts the

P15, l19: The sentence is confusing. Please revise.

The line has been reworded to read, "The IQRSS values show that the prototype tends to produce sharper forecasts than HE early in the season i.e. for forecasts initialised in January and February in cluster S^1 and forecasts initialised in January, February and March in clusters S^2 and S^3. This is reversed for the remaining initialisation dates where HE tends to produce sharper forecasts than the prototype."

P6, l10: "subbasins sub-basins"

The second hyphenated instance has been deleted.

Response to reviewer's comments in SC1

Reviewer's comments are in blue, our comments are in black.

Summary of the manuscript

The manuscript shows, how a hydrological seasonal forecast system prototype is adjusted from the previous version and evaluated on its ability to predict spring flood volumes in Swedish rivers. The aim is to improve water resource management for hydropower decision makers. The study area consists of 84 subbasins in northern Sweden, which have a runoff regime that is strongly influenced by spring snow melt. The skill of the multi- model prototype is compared in cross-validated hindcasts to the historical ensemble streamflow prediction based on measurements between 1981 and 2015. This historical ensemble represents the setup currently used for hydropower reservoir management. The multi-model prototype represents combinations of the historical ensemble, an analogue ensemble (subset of the historical ensemble based on similarities in parameters of interannual climate variability), a dynamic modelling ensemble (bias-corrected season- al forecast) and a statistical modelling ensemble (downscaled seasonal forecast). Several complementary, statistical measures were used for the evaluation of the new proto- type. The prototypes, that combine 3 different ensembles show at best a significant improvement compared to the currently used historical ensemble and at worst a comparable skill.

Main assessment

Based on our assessment, the reviewed manuscript reaches a substantial conclusion based on sufficient results which generally were outlined clearly and used valid assumptions. Overall, we like the clear structure of the paper and the scientific notation. In the abstract and the introduction, it is nicely explained why there is a public benefit behind this research.

When we first read the introduction, we had some problems to understand the differences between the two approaches (dynamical and statistical). When we came to the points where it is explained better it is not a problem anymore. Maybe a quick hint to the section 2.14 and 2.15 would help the readers - or provide some of the clarifications already in the introduction.

1)

We have added a cross reference to page 2, line 11 and line 12. It now reads as follows:

"In practice, there are two predominant approaches to making hydrological forecasts at the seasonal scale; statistical approaches and dynamical approaches (see Sect. 2.1.4 and Sect. 2.1.5 for more regarding these approaches in the context of this work)."

It would have been useful to add the research questions to the introduction. The paper describes what has been done, how it was tested and how good the new prototype is for catchments in Sweden but not explicitly what science question or hypothesis is answered.

2)

This is a valid point. We have added a sentence after page 1, line 13 and line 14 to include what the overarching hypothesis of this work is. It reads:

"The hypothesis explored in this work is that a multi-model seasonal forecast system which incorporates different modelling approaches is generally more skilful at forecasting the SFV in snow dominated regions than a forecast system that utilises only one approach."

This, together with the beginning of the paragraph, now gives the reader a brief description of the issue this work is intended to address and the hypothesis that is tested.

The paper also does not discuss how applicable this method/prototype is for other snow dominated areas.

3)

We have added a paragraph that discusses this to the results and discussion section. See our response (no.7) to your comments:

Thanks to the SMHI, the datasets are very good, but we don't understand why the subbasins with 61, 63, and 73 percent of missing data are also included in these 84 basins that are investigated. Has it at least been checked whether and how the results change when those three subbasins are excluded?

4)

These stations were included because they are part of the operational forecast and are therefore relevant to the prototype. This we mention on page 11, line 25 and line 26. We did check how their inclusion affected the results and found that there was a small improvement in the skill of the prototype. This we attribute to an adverse affect they have on the statistical branch which needs to be trained on these historical data. However, the drop in skill is not such that it detracts from the general results. So, due to

their importance to the prototype and the relative low impact hey have on the performance we chose to retain them in this work.

The evaluation part from this study is very elaborate and we like this. However, it could have been more clearly explained what a 6% improvement means, for example in volume of water or economic value for hydropower generation. Is this 6% really significant?

5)

Please see our response (4^{th}) to reviewer 2. Yes it is significant to the power companies. It is difficult to put clear contextual examples of what this improvements means economically.

We have made changes to the first third of the conclusion section (see our response no. 41 to your comments). In this we have included a volumetric interpretation of what a 6% reduction in SFV entails. The bullet point reads:

"• The prototype is able to reduce the forecast error by 6% on average. This translates to an average volume of 9.5 x 10^6 m^3."

We would have appreciated it if the paper provided a little more information about the seasonal meteorological forecasts that are used. For example something about the uncertainty of these models or why the ECMWF IFS system is used (is there no other meteorological forecast system for six months or is it the best seasonal forecast system)?

6)

The choice to use the ECMWF as a provider of the seasonal forecasts to force the prototype is primarily based on operational considerations. SMHI has operational access to these products so no extra effort is needed to source and collect these data. It should be noted that there are other seasonal forecast data providers and we are looking to test them in the future, however that is not part of the scope of this work.

We have added two sentences after the line on page 11, line 20 briefly motivating our choice. They read:

"The choice to use ECMWF data is primarily a practical one. The ECMWF is an established and proven producer of medium range forecasts and SMHI already has operational access to their products."

Finally, it would have been interesting if the differences for the different catchments would have been discussed. Were the improvements mainly seen for large/small catchments, for high elevation/flat catchments? Some maps would have been nice as well.

7)

We have added a section to the Results and discussion section which addresses this point. It reads:

"3.4 Spatial and temporal variations and transferability of the prototype

Both multi-model ensembles show skill at forecasting SFV with respect to forecast error, ability to reproduce the interannual variability in SFV, and the ability to discriminate between BN, NN, and AN events. The prototype, in particular, is at worst comparable to the HE and at best clearly more skilful. This relative performance of the prototype varies both in space and time. Figure 7 shows maps of the median bootstrapped FY+ values. For hindcasts initialised in January the spatial pattern in the FY+ scores show that the prototype tends to outperform HE more in subbasins that have a higher latitude or elevation. However, as the initialisation date approaches the spring flood period this pattern becomes less and less coherent. This general pattern is also true for MAESS scores. This suggests that the change in the performances of the prototype and HE, as a function of initialisation date, are not always similar for subbasins that are near one another. Further work would be needed to find out what the underlying reason for this is.

Data availability is the biggest limiting factor to the transferability of this approach to other areas. The HE, AE, and SE approaches are all dependant on good quality observation time-series. Additionally, the skill all three of these approaches would be expected to be affected by length of these time-series. They length of the time-series should be long enough to be a good representative sample of the climatology otherwise the forecasts would be biased in favour of the climate represented in the data and not the true climatology.

The SE and AE approaches require an understanding of how the variability in the local hydrology is affected by large scale circulation phenomena such as teleconnection patterns to help select predictors and teleconnection indices for inputs to each approach respectively. The hydrological rainfall-runoff model used in the prototype should not pose a problem, although HBV has been successfully setup for

snow dominated catchments outside of Sweden (e.g. Seibert et al., 2010; Okkonen and Kløve, 2011), any sufficiently well calibrated rainfall-runoff model would suffice.

We believe that, if the above requirements are met, a seasonal hydrological forecast system similar to the prototype can be setup in other snow dominated regions around the world."

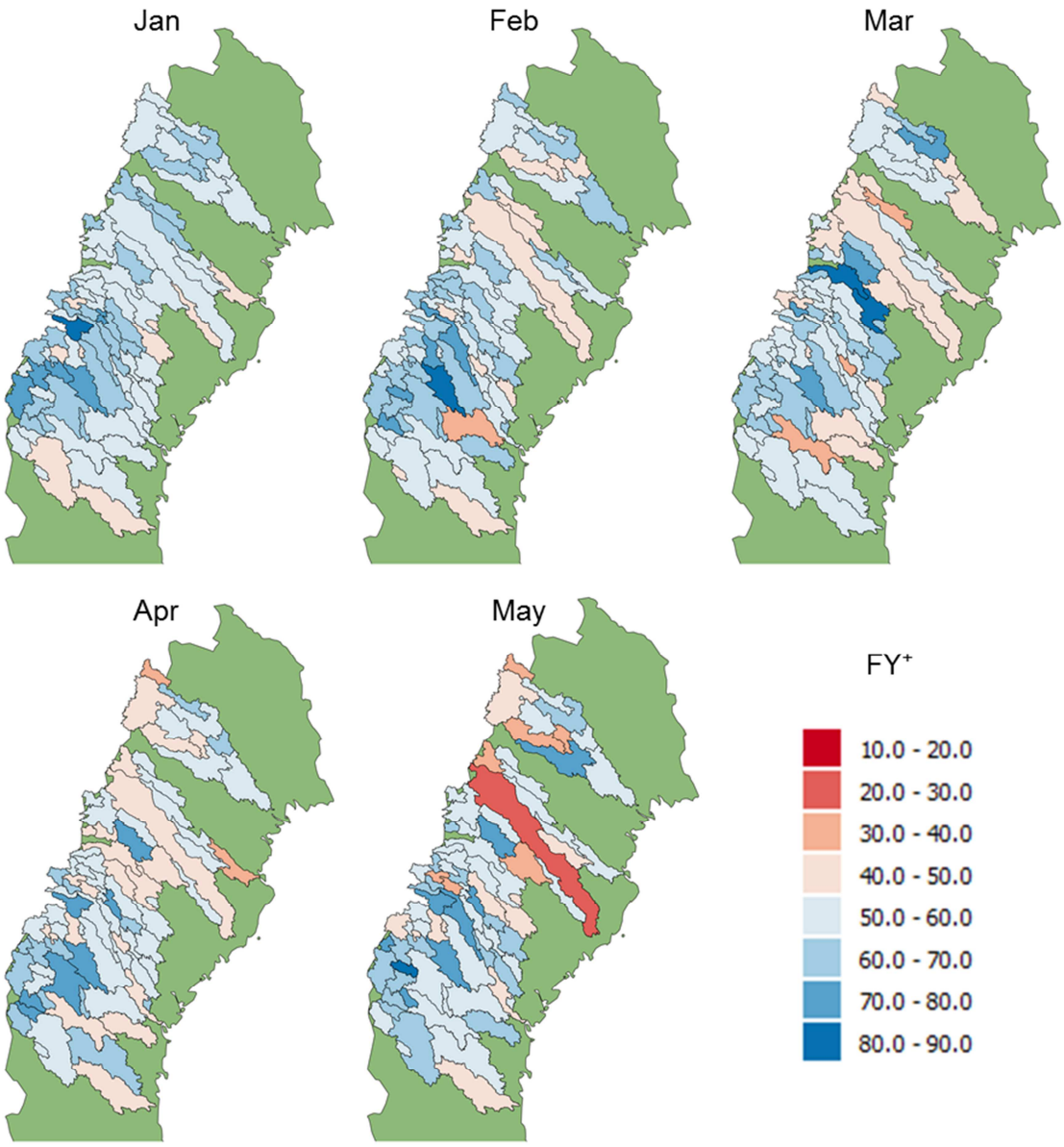

Figure 7. Maps of the median bootstrapped FY+ values for each of the initialisation dates.

List of major and minor points

Page one:

Lines 8-10: In our opinion a very catchy and smart opening.

8)

Thank you.

Line 15 to 18: This sentence is maybe too long and little too complicated for the abstract. (Full stop before ‘however’?). Twice 'considered'

9)

This sentence has been broken up and now reads:

“Both the multi-model methods considered showed skill over the reference forecasts. The version that combined the historical modelling chain, dynamical modelling chain, and statistical modelling chain performed better than the other and was chosen for the prototype.”

Line 23: Unclear reference, presumably ‘Statistiska centralbyrån’? For clarity, the abbre- viation can be included in the reference (Statistiska centralbyrån (SCB): …)

10)

The reference in page 1, line 23 has been changed from the abbreviation to the full reference.

Line 27 to 28: The idea or point behind the sentence comes across. But if you first read this sentence, it could be puzzling. We also think that with all these brackets the text looks not as nice as it could. Why not just add 'and vice versa' at the end of the sentence?

11)

Yes the use of vice versa does improve the readability of the sentence without detracting from the message. The sentence has been rewritten as suggested. It now reads:

“This reservoir management is important as the energy demand is out of phase with the natural availability of the water resources; typically demand is higher during the colder months when the inflows are lower and vice versa.”

Page two:

Line 3: Grammatical: The strategy is to have reservoirs which are then managed. Line 4: comma: To achieve this, operators …

12)

Please see our response (4th) to reviewer 1.

Line 6: The meaning of the expression ‘sources of predictability’ is unclear to us in this situation. Can you explain briefly?

13)

The expression refers to where the signal that gives skill to the forecasts originates from. The SFV is a function of many hydrometeorological factors but some influence the variability of the SFV more than others. For example, in the context of this work, the snowpack is a major contributor to the SFV and therefor data related the amount of water stored in the snowpack can potentially be used to make a skilful forecast. In this example, information regarding the snow pack is leading source of predictability in seasonal forecasts of the SFV in these regions.

Line 7: Decide on stores within or in the catchment. We suggest using within. Line 11: Probably : instead of ; after ‘forecasts at the seasonal scale’.

14)

Yes, the former is a typo and the redundant word ‘in’ has been deleted. The latter suggestion of using a colon instead of a semi-colon has also been applied.

Line 15: There is no need for a comma after the closed bracket. Both times. Lines 19, 20: The second time ‘force it’ (end of line 19) is not necessary.

15)

These changes have been applied.

Lines 31, 32: Do the references explain how and to what extent historical observations of precipitation and temperature are possible representations of future meteorological conditions in the context of the ongoing climate change? Do the historical data show any significant trend?

16)

No they do not. The standard ESP approach assumes stationarity and does therefore not take into account changes in climate. Yes, there is a change signal in the historical data but making allowances for this was not within the scope of this work. However, there are future plans to investigate the added value of adjusting the historical data to mitigate this change signal before use in the modelling chain.

Another approach, which we mention in the manuscript, is the post-processing of the forecasts to account for any biases related to factors such as climate change signals.

Page three:

Line 13: Maybe already mention here the number of catchments and data period. That way we already know something about how these modeling steps are applied. Other- wise the 35 in Line 35 is not so clear.

17)

We feel that by adding this information in page 3, line 13 would make the sentence clumsy to read. Instead, we added information on the number of catchments to page 3, line 21; we added information regarding the data period to page 3, line 30. The affected sentences now read:

"The aim is to adapt their methodology for use in an operational environment and then evaluate the resulting prototype against the current operational system using cross-validated hindcasts for 84 gauging stations in northern Sweden (see sect. 2.6)."

And

"These outputs are pooled together rather than using an asymmetric weighting scheme due to the lack of data points, a total of 35 spring flood events (hindcast period was 1981-2015, see Sect. 2.6), from which to derive a robust weighting scheme."

Line 14 to 15: Twice the word "brief"

18)

The second occurrence of brief has been removed

Line 19: first improved by Foster et al. (2010) and, later improved upon and first tested by Olsson et al. (2016).

19)

A comma was added.

Lines 23, 24: Here the manuscript includes already some results but is in the Materials and Methods section. Finish sentence after '… of these four were tested.'

20)

The sentence was shortened accordingly.

Line 28: Replace ‘relevant’ by ‘respective’.

21)

The replacement was made.

Page four:

Line 7: Is there a reference on what the seasonal forecasting practice at SMHI is? Line 9: It is not clear to us how the DBS method is different from the previous method.

22)

We assume you are referring to page 5, line 7. Please see our 3rd response to reviewer 2.

Line 26: We are not familiar with the teleconnection approach. A brief explanation would have been useful.

23)

There is a brief explanation of the revised teleconnection approach later in the next paragraph (page 4, line 31 – page 5, line 5) which gives an overview of what the approach entails. However, we have changed the word ‘the’ to ‘their’ (page 4, line 26) to clarify that we are referring to an approach proposed by Olsson et al. (2016) which we have already referred to.

Page five:

Line 2: What is the meaning of and the justification for a distance of 0.2?

24)

We are using the persistence in the teleconnection indices leading up to the forecast date to select analogue years out of the historical dataset. In order to be able to identify which of the historical years are analogues we need a selection criteria. We define an analogue to be any year whose Euclidean distance is less than 0.2 units from the Euclidean position of the ‘current’ year (see page 5, line1 – line2). The threshold is a compromise between being small enough to be sufficiently specific and being large enough to actually be able to capture some analogues from the historical data.

We appreciate that this is not entirely clear in the manuscript. We added to page 5, line 2 to clarify what the value 0.2 referred to. It now reads:

"If the values of these indices are considered to be coordinates in Euclidean space we defined analogue years to be those years whose positions are within a distance of 0.2 units in the Euclidean space from the position of the forecast year."

Additionally, we have added a line directly after the sentence in question giving further information as discussed above. This line reads as follows:

"The threshold is a compromise between being small enough to ensure that the climate setup is indeed similar to the year in question and being large enough to actually be able to identify some analogues from the historical ensemble."

Line 29 and following: Is there a reference (needed) for the physical support of the asymmetric weighting?

25)

The sentences in the manuscript directly following this line (page 5, line 29 – page 6, line 2) give our account for this physical support. For example, we explain that the relative importance of the snowpack earlier in the season is less than it is later in the season with respect to the coming meteorological conditions.

Page six:

Line 24: Typo: overfitted

26)

We have corrected the typo

Line 28: So n = 35 in this case?

27)

Yes, we have clarified this by changing the sentence to now read:

"This process is repeated n times to give a validation dataset of length n, for this work n=35."

Page seven:

Line 27: Why is the relative mean absolute error used (error divided by SFV0y)? Would the absolute error (without the division) not be a stronger focus on the flood peaks and therefore more beneficial for the assessment of the skill of forecasting the spring flood? Please explain.

28)

The division operation converts the error form a volume to a ratio of the observed volume. This does not alter the relative emphasis of the metric, but it does make it more intuitive.

Line 28: The superscripts suggest that it is SFV to the power of y. A different notation is clearer.

29)

Another notation would be clearer from a mathematical understanding, however this notation is fairly common in the hydrology literature. Additionally, by retaining the current notation we are maintaining continuity with the previous works which this work builds on.

Page ten:

Line 6: Subbasin.

30)

Please see our response (5^{th}) to reviewer 1.

Line 6: There should be a reference to figure 3 included in the sentence.

31)

We added a cross reference to figure 3 at the end of the sentence.

Line 10: How is total runoff divided between the 3 subbasins? Would it be interesting to use the new setup also for the two other subbasins?

32)

The three clusters S1, S2, and S3 are made up of subbasins from seven river systems. They are not subbasins in themselves. Table 2 gives some basic statistics regarding the SFV in the subbasins of the different clusters. The prototype is aimed primarily at reservoir operators in the hydropower industry and the majority of the large operations are based in these three clusters. We agree that it would be interesting to apply this approach to the other two clusters but this will have to wait for now.

Line 17: SFV already introduced on page 3.

33)

Please see our response (6^{th}) to reviewer 1.

Line 27: We don’t understand why the subbasins with 61, 63, and 73 percent of missing data are also included in the 84 basins that are investigated. Has it been checked whether and how the results change when excluding those three subbasins?

34)

The inclusion of these basins is due to them being part of the current operational forecast system (see page 10, line 29 – line 29) and will be required going forward. We checked how their inclusion affected the results. Their inclusion typically reduced the apparent added value of the prototype over the current operational forecast system due to the need to use this data to train the statistical model. However, the reductions in the skill scores were not statistically significant.

Line 31: What is the PTHBV dataset from SMHI? is this the ptq file for HBV (but without the q)?

35)

Yes. PTHBV is the name of a gridded product for Sweden of P and T observations which is used to populate the ptqw file in HBV. The Q and W values are populated using station data.

Page eleven:

Line 4: Do you refer to the performance measures for the HBV rainfall-runoff model of the historical data?

36)

Yes, these are validation scores for HBV using perfect hindcasts i.e. forced using observed P and T.

Line 7: The ‘than’ near the end of the line is missing one or two adjectives describing lati- tude and elevation. Decide on either and or or.

37)

We have reworded this line to now read:

“Subbasins in cluster S1 are typically at a latitude or elevation lower than those in clusters S2 and S3, similarly the subbasins in S2 with respect to those in S3.”

Line 22-24: We don't really understand where the 15 and 51 come from. We probably missed that somewhere, or is it not explained anywhere?

38)

These are the number of ensemble members that are available in the seasonal forecasts/hindcasts from the ECMWF. We have reworded page 11, line 23 to make this clearer in the context of the surrounding paragraph. It now reads:

"This is because the number of ensemble members available in the ECMWF seasonal forecast is limited to 15 for the hindcast period while the operational seasonal forecast ensemble has 51 members."

Page thirteen:

Line 21: Are these 5% of the catchments located in a certain area or do they have similar characteristics? Large/small, steep/flat, northern/southern?

39)

As we are more interested in the overall performance we did not put an emphasis on decomposing the possible reasons behind why the results for specific stations may have performed the way they did. Your questions are valid and should be investigated going forward, but in the context of this work it is less important. However, results from both this work and previous work by Olsson et al. (2016) suggest that the subbasins where the prototype does not perform as well tend to be located in, but not isolated to, the middle and lower reaches of the rivers. Also, please see our earlier response (no. 7).

Page fourteen:

Line 29: Why was exactly this site choose for the analysis of the forecast ensemble sharpness? Just one sentence, why exactly this basin is relevant out of the whole set of the 84 sites.

40)

The following sentence was included after page 14, line 29. It reads:

"This basin was chosen as an example of a where the prototype showed typical performance results i.e. neither the best nor the worst."

Page sixteen:

Line 5-18: This is more discussion than real conclusion. Move?

41)

We moved the parts that were more discussion in nature to the results and discussion section. This section now reads:

“In this paper we present the development and evaluation of a hydrological seasonal forecast system prototype for predicting the SFV in Swedish rivers. Initially, two versions of the prototype, MEads and MEhds, were evaluated together with the HE using climatology as a reference to both help select which version of the prototype to proceed with and to get a general impression of their skill to forecast the SFV. Thereafter the chosen prototype was evaluated using HE as a reference and finally the sharpness of the hindcast ensembles were analysed.

The main findings are summarized below:

- The prototype is able to outperform the HE approach 57% of the time on average. It is at worst comparable to the HE in forecast skill and at best clearly more skilful.
- The prototype is able to reduce the forecast error by 6% on average. This translates to an average volume of 9.5 x 106 m3.
- The prototype is generally more sensitive to uncertainty, that is to say that the ensemble spread tends to be more correlated with the forecast error. This is potentially useful to users as the ensemble spread could be used as a measure of the forecast quality.
- The prototype is able to improve the prediction of above and below normal events early in the season.”

Figures:

Figure 2: In the figure description it is mentioned that the spring flood is in the period be- tween the onset and the last day of July. On the x-axis the days since the 1st January are written. We think that for the reader it would be easier if on the x-axis the first of the months were labelled and the date of the 80 days threshold was indicated (20th or 21st of March).

42)

We disagree with the need to change the x-axes labels to dates. We feel that by doing so would complicate the figure more than it simplifies it. Figure 2 is meant as a generalised schematic showing the concept of how we define the spring flood period and not meant to give specific dates to the readers which are not mentioned elsewhere in the manuscript. The only deviation from what is mentioned in the body of the manuscript is that the 31st July is not explicitly.

As a compromise we have added the day numbers for the 31st July in parentheses the date is mentioned in the figure description. It now reads:

“The spring flood is the period between the onset and the last day of July (day 211/212 since the 1st January).”

5

The development and evaluation of a hydrological seasonal forecast system prototype for predicting spring flood volumes in Swedish rivers

Kean Foster[1, 2], Cintia Bertacchi Uvo[2], Jonas Olsson[1]

[1]Research & Development (hydrology), Swedish Meteorological and Hydrological Institute, 601 76 Norrköping, Sweden
[2]Department of Water Resources Engineering, Lund University, Box 118, 221 00 Lund, Sweden

Correspondence to: Kean Foster (kean.foster@smhi.se)

Abstract. Hydropower makes up nearly half of Sweden's electrical energy production. However, the distribution of the water resources is not aligned with demand, most of the inflows to the reservoirs occur during the spring flood period. This means that carefully planned reservoir management is required to help redistribute the water resources to ensure optimal production and accurate forecasts of the spring flood volume (SFV) is essential for this. The current operational SFV forecasts use a historical ensemble approach where the HBV model is forced with historical observations of precipitation and temperature. In this work we develop and test a multi-model prototype, building on previous work, and evaluate its ability to forecast the SFV in 84 sub-basins in northern Sweden. The hypothesis explored in this work is that a multi-model seasonal forecast system incorporating different modelling approaches is generally more skilful at forecasting the SFV in snow dominated regions than a forecast system that utilises only one approach. The testing is done using cross-validated hindcasts for the period 1981-2015 and the results are evaluated against both climatology and the current system to determine skill. Both the multi-model methods considered showed skill over the reference forecasts. The version that combined the historical modelling chain, dynamical modelling chain, and statistical modelling chain performed better than the other and was chosen for the prototype. The prototype was able to outperform the current operational system on average 57% of the time and reduce the error in the SFV by ~6% across all subbasins and forecast dates.

1 Introduction

The spring flood period (sometimes referred to as the spring melt or freshet period in the literature) is of great importance in snow dominated regions like Sweden where hydropower accounts for nearly half of the country's electrical energy production (Statistiska centralbyrån, 2016). Between 55-70% of the annual inflows to reservoirs in the larger hydropower producing rivers occur during this relatively short period, typically from mid-April/early-May to the end of July. This means that the majority of the annual water resources available for hydropower production would only be available to producers during this period if it were not regulated through carefully planned reservoir management. This reservoir management is important as the energy demand is out of phase with the natural availability of the water resources; typically demand is higher during the colder months when the inflows are lower and vice versa. Therefore the goal is to redistribute the

availability of these resources from the spring flood period to other times of the year when electricity demand is higher i.e. during the colder winter half year, while maintaining a balance between a sufficiently large volume of water for optimal production and enough remaining capacity for safe flood risk management (Olsson et al., 2016). The typical strategy for operators in Sweden is to have reservoirs at around 90% capacity at the end of the spring flood which is then ideally maintained until the beginning of winter. To achieve this, operators require reliable seasonal forecast information to help them in planning the operations both leading up to and during the spring flood period.

The sources of predictability for hydrological seasonal forecasts come from the initial hydrological conditions i.e. information relating to the water stores within the catchment (e.g. Wood and Lettenmaier, 2008; Wood et al. 2015; Yossef et al., 2013), and also from knowledge of the weather during the forecast period i.e. seasonal meteorological forecasts (e.g. Bennet et al., 2016; Doublas-Reyes et al., 2103; Wood et al. 2015; Yossef et al., 2013). Hydrological seasonal forecasts attempt to leverage at least one of these sources of predictability to make skilful predictions of future streamflow.

In practice, there are two predominant approaches to making hydrological forecasts at the seasonal scale, statistical approaches and dynamical approaches (see Sect. 2.1.4 and Sect. 2.1.5 for more regarding how these approaches in the context of this work). Statistical approaches utilise empirical relationships between predictors and a predictand, typically streamflow or a derivative thereof (e.g. Garen, 1992; Pagano et al., 2009). These predictors can vary greatly in type from local hydrological storage variables like snow and groundwater storages (e.g. Robertson et al. 2013; Rosenberg et al., 2011), to local and regional meteorological variables (e.g. Còrdoba-Machado et al., 2016; Olsson et al., 2016), to large scale climate data such as ENSO indices (e.g. Schepen et al., 2016; Shamir, 2017). All, however, are trying to leverage the predictability in these predictors that originate from one of the two aforementioned sources. Dynamical approaches use a hydrological model, typically initialised with observed data up to the forecast date so that the model state is a reasonable approximation of the initial hydrological conditions, and then force it with either historical observations (called ensemble streamflow prediction or ESP; e.g. Day, 1985) or force it using data representative of the future meteorological conditions such as general circulation model (GCM) outputs (e.g. Crochemore et al. 2016; Olsson et al. 2016; Yuan et al. 2013, 2015, 2016). Attempts to improve these types of approaches have involved bias adjusting the GCM outputs (e.g. Crochemore et al., 2016, Lucatero et al., 2017; Wood et al., 2002; Yuan et al. 2015) or bias adjusting the hydrological model outputs (e.g. Lucatero et al., 2017) or a combination of both (e.g. Yuan and Wood, 2012). Another dynamical approach is the well-established ESP method (Day, 1985). This is similar to the previous approach, however instead of using GCM outputs to force the hydrological model it uses an ensemble of historical data. This approach is perhaps one of the most widely used methods and is still the subject of new research. Recent work have looked at conditioning the ensembles before using them, this conditioning can be done using GCM outputs (e.g. Crochemore et al., 2016), climate indices, and circulation pattern analysis (e.g. Beckers et al., 2016; Olsson et al. 2016; Yossef et al., 2016).

The current practice at the Swedish Meteorological and Hydrological Institute (SMHI) for seasonal forecasts of reservoir inflows is the ESP approach. It assumes that historical observations of precipitation and temperature are possible representations of future meteorological conditions and are used to force the HBV hydrological model (e.g. Bergström,

1976; Lindström et al., 1997) to give an ensemble forecast that has a climatological evolution from the initial conditions. A number of attempts have been made in the past to improve the performance of these spring flood forecasts with limited success (Arheimer et al., 2010) demonstrating that these seasonal forecasts are already of a high quality. Work by Olsson et al. (2016) on improving these forecasts was able to realise reasonable improvements using a multi-model approach. By combining a statistical approach, dynamical approach and an analogue approach (conditioned ESP) they were able to show a ~4% reduction in the forecast error of the spring flood volume (SFV). The purpose of this paper is to continue on and update the work started by Olsson et al. (2016) to develop and evaluate a hydrological seasonal forecast system prototype for forecasting the spring flood volumes in Sweden.

This paper is organised as follows. Section 2 outlines the prototype, including the individual model chains, the experimental setup, the methods and tools used, and the study area and data used in this work. Section 3 presents and discusses the cross-validated evaluation scores for the prototype, first with reference to climatology and then with reference to the current operational system that is in use at SMHI. Section 4 concludes with the main findings and a brief outlook for future work.

2 Materials and Methods

2.1 The multi-model system and the individual modelling chains

In this section we present the modelling approaches used in this work. These are based on those explored by Olsson et al. (2016) with some modification to facilitate their use in an operational environment. First we briefly present the multi-model prototype (Sect. 2.1.1) followed by a brief overview of the individual modelling chains used in the multi-model and why we chose them (Sect. 2.1.2 – 2.1.5). For more information regarding the individual modelling chains readers are referred to Olsson et al. (2016) and the accompanying supplement.

2.1.1 The multi-model ensemble (ME)

The prototypes developed in this work builds on an approach first proposed by Foster et al. (2010), and later improved upon and first tested by Olsson et al. (2016). The aim is to adapt their methodology for use in an operational environment and then evaluate the resulting prototype against the current operational system using cross-validated hindcasts for 84 gauging stations in northern Sweden (see Sect. 2.6). Four different modelling chains were considered when developing the prototype (Sect. 2.1.2 – 2.1.5). The performances of different combinations of these four were tested.. A combination of all four modelling chains was not considered as the analogue model chain is a subset of the historical model chain.

Figure 1 shows the generalised schematic of the two prototypes, ME_{ads} and ME_{hds} (where the subscripts refer to the individual modelling chains making up each multi-model), including where the current methodologies differ significantly from those in previous works. These differences are discussed in the relevant modelling chain sections below. The prototypes are multi-model ensembles of the outputs from the three respective individual modelling chains. These outputs are pooled together rather than using an asymmetric weighting scheme due to the lack of data points, a total of 35 spring

flood events (hindcast period was 1981-2015, see Sect. 2.6), from which to derive a robust weighting scheme. The simple weighting scheme used by Olsson et al. (2016) was tested but, other than improving the ensemble sharpness, did not offer an improvement over the pooling approach.

2.1.2 Historical ensemble (HE)

The historical model chain, the dark blue chain third from the left in Figure 1, is an ensemble forecast made by forcing a rainfall-runoff model with historical observations of precipitation and temperature. This approach is often referred to as ESP (ensemble streamflow prediction) in the literature but we chose not to as we feel our terminology is more descriptive in the context of this work. This is the current operational seasonal forecasting practice at SMHI. The HBV model (Bergström, 1976; Lindström et al. 1997) is initialized by using observed meteorological inputs (P and T) to force the model up to the forecast date so that the model state reflects the current hydro-meteorological conditions. Then, typically all available historical daily P and T series for the period from the forecast issue date to the end of the forecasting period are used as input to HBV, generating an ensemble of forecasts that are climatological in their evolution from the initial conditions. The HE is used as the reference ensemble unless otherwise stated.

The HBV is run one river system at a time and the model outputs are later regrouped into three clusters (Section 2.6). Typically only historical data prior to the forecast date are used to force the model, however to allow for a more robust cross validation all data including for years after the forecast date were used (excluding the year in question of course). Unfortunately, the scope of this work did not allow for the recalibration of the HBV model before each cross validated hindcast. This will potentially inflate the performance of the model for the hindcasts of years that were used in the calibration of the model. This will affect the analogue and dynamical model chains too as they also incorporate the HBV model in their setup.

2.1.3 Analogue ensemble (AE)

The analogue model chain, the light blue chain furthest to the left in Figure 1, is a subset of the HE. The hypothesis is that it is possible to identify a reduced set of historical years (an analogue ensemble) that describes the weather in the coming forecasting period better than the full historical ensemble used in HE. In this work the circulation pattern approach used by Olsson et al. (2016) was omitted due to data availability issues making it impractical for operational applications. Additionally, their teleconnection approach was revised to take advantage of the findings by Foster et al. (2012) and Foster et al. (2016) where they identified which teleconnection patterns are related to the SFV and for which period of their persistence prior to the spring flood this connection is strongest. The teleconnection indices they identified are the Arctic oscillation (AO) and the Scandinavian pattern (SCA) and the periods of persistence for these indices, expressed as the index mean for the identified period, are the seven and eight months leading up to the spring flood respectively..

The persistence for each teleconnection index is calculated from the beginning of the aforementioned period to one month prior to the forecast date (a limitation imposed by data availability), similarly this was done for all years in the climatological ensemble. If the values of these indices are considered to be coordinates in Euclidean space we define analogue years to be those years whose positions are within a distance of 0.2 units in the Euclidean space from the position of the forecast year. The threshold is a compromise between being small enough ensuring that the climate setup is indeed similar to the year in question and being large enough to actually be able to capture some analogues from the historical ensemble.The selection of the analogues is done at the regional scale, by cluster (Section 2.6), and these selections applied to the associated sub-catchments in turn.

Similar to the HE, the analogue method makes use of both prior and later years to the hindcast year for the cross validated hindcasts.

2.1.4 Dynamical modelling ensemble (DE)

The dynamical model chain, the dark red chain furthest to the left in Figure 1, is similar to the HE; an adequately initialised HBV model is forced by an ensemble of seasonal forecasts of daily P and T from the ECMWF IFS system 4 (Sect. 2.7 Seasonal data). A change to previous work has these daily P and T data bias adjusted first before being used to force HBV. The bias adjustment method used is version of the distribution based scaling approach (DBS; Yang et al., 2010) which has been adapted for use on seasonal forecast data. DBS is a quantile mapping bias adjustment method where meteorological variables are fitted to appropriate parametric distributions (e.g. Berg et al., 2015; Yang et al., 2010). For precipitation, two discrete gamma distributions are used to adjust the daily seasonal forecast values, one for low-intensity precipitation events ($\leq$ 95th percentile) and another for extreme events (> 95th percentile). For temperature, a Gaussian distribution is used to adjust the daily seasonal forecast values.

Observed (Sect. 2.6 Study area and local data) and seasonal forecast (Sect. 2.7 Driving Data) time-series of P and T spanning the relevant forecast timeframe (e.g. Jan-Jul for forecasts initialised in January) and for the reference period 1981-2010 are used to derive the adjustment factors to transform the seasonal forecast data to match the observed frequency distributions. First the precipitation data is adjusted then the temperature data. The latter is done separately for dry and wet days in an attempt to preserve the dependence between P and T (e.g. Olsson et al. 2010; Yang et al, 2010). Adjustment factors are calculated for each calendar month as the distributions can have different shapes depending on the physical characteristics of the precipitation processes that are dominant. It should be emphasized that the adjustment parameters were estimated using much of the same data to which they were applied. Ideally the parameters would be estimated using data that does not overlap the data which is being adjusted. However, this was not possible in the scope of this work.

There have been some criticisms raised lately regarding the applicability of quantile mapping for bias adjusting seasonal data (e.g. Zhao et al., 2017). They point out that although quantile mapping approaches are effective at bias correction they cannot ensure reliability in forecast ensemble spread or guarantee coherence. Unfortunately, the scope of this work did not

allow the testing of other bias adjustment methods but the criticism is noted and further work is planned to address these points.

These bias adjusted data are then converted into HBV inputs by mapping them from their native grid onto the HBV sub-catchments. The mapping is done by areal weighting and the resulting sub-catchment average P and T values are then adjusted to represent different altitude fractions within the catchment. These data are then used to force the HBV model from the same initial state as used in the HE procedure.

No changes to this methodology are needed to accommodate the cross-validated hindcasting as done with the other model chains.

2.1.5 Statistical modelling ensemble (SE)

The statistical model chain, the orange chain second from the left in Figure 1, is an ensemble forecast produced by downscaling forecasted or modelled large-scale variables (predictors) to the SFV for each cluster (predictand). The downscaling is done using an SVD approach (singular variable decomposition). The predictors are three large scale circulation variables (Section 2.7) and the modelled snow depths from the HBV initial conditions. The outputted ensembles of SFV are combined using a simple arithmetic weighting system. The normalised squared covariance between the four predictors and the predictands are ranked for each forecast initialisation date and weights between 0 and 1 are applied to the different predictors according to their rank. The lowest ranked predictor is assigned a weight of 0.1 (= 1/10), the next lowest predictor is assigned a weight of 0.2 (= 2/10), and so on until all four have been assigned a weight. The reason that an asymmetric weighting scheme is used here is that there is physical support for it. Early in the season the snowpack, which is the majority contributor to the spring flood volume, is still a fraction of what it will be and is still accumulating. Therefore, the coming meteorological conditions, which dictate snowpack evolution, are more important earlier on in the season than they are later giving physical support for asymmetric weighting. Additionally, the relative importance of these meteorological predictors with respect to each other differs with time too.

The relative simplicity of the statistical model chain means that it was possible to retrain the model before each hindcast during the cross-validation calculations allowing for no overlap between the calibration and validation periods.

2.2 Defining the spring flood

In previous works the spring flood period has often been defined in terms of calendar months e.g. May-June-July (Nilsson et al., 2008; Foster and Uvo, 2010; Arheimer et al., 2011; Olsson et al., 2016; Foster et al., 2016). This definition of the spring flood period is not ideal as it does not take into account the interannual and geographical variations in the timing of the spring flood onset. In this work we propose an improvement to this practice where we define the spring flood to be the period from the onset date to the end of July.

We define the onset as the nearest local minima in the hydrograph before the date after which the inflows are above the 90th percentile, with reference to the inflows during the first 80 days of the current year, for a period of at least 30 days (Figure 2). For forecasts made after January i.e. those made in February, March, April, and May, the missing inflow data between the 1 January and the forecast date are filled with simulated inflow data from the HBV model using observed precipitation and temperatures as input data.

A drawback to this definition is that it is not comprehensive as the end of the spring flood is not defined according to the hydrograph but rather by date. The reasons for not defining the end of the spring flood objectively are twofold. Firstly, the forecast horizon for the ECMWF-IFS is seven months which means that forecasts initialised in January may not encompass the entire spring flood period, and secondly, a robust and objective definition of what constitutes the end of the spring flood was difficult to realise within the scope of this work. Further work is needed to resolve this in a more satisfactory manner.

2.3 Experimental setup

The challenge in this work was to perform a robust evaluation on a limited dataset (35 spring floods, 1981-2015) while minimising the risk of unstable or over fitted statistics. Therefore, a leave-one-out cross validation (LOOCV) protocol was adopted. Additionally, as it was not practical to recalibrate the HBV model before each step of the LOOCV process; the statistical model uses the same periods for training as those used to calibrate and validate the HBV model. LOOCV is a model evaluation technique that uses n-1 data points to train the models and the data point left out is used for validation. This process is repeated n times to give a validation dataset of length n. This allowed for a more robust evaluation with a limited dataset and to be able to sample more of the variability in the training period than if a traditional validation were performed. The second point is especially advantageous for evaluating the statistical model which is especially sensitive to situations that were not found within the training period. LOOCV was applied to the individual model chains.

To assess the relative skill for different lead times, we evaluate hindcasts issued on the 1st of January (Jan), 1st of February (Feb), 1st of March (Mar), 1st of April (Apr), and 1st of May (May) for the spring floods 1981-2015. The evaluation of performance is done in terms of how well the SFV is forecasted.

2.4 Evaluation

As it has been mentioned above, we are interested in the ability of a multi-model ensemble's ability to forecast the SFV at differing lead-times i.e. forecasts initialised on the first of the month for the months of January through May. It was suggested by Cloke and Pappenberger (2008) that for a rigorous assessment of the quality of a hydrological ensemble prediction system (HEPS) it is not only important to select appropriate verification measures but also to use several different measures so that different properties of the forecast skill can be estimated resulting in a more comprehensive evaluation.

The evaluations in this paper are designed to answer the following questions:

- Can the forecasts improve on the reference forecast error?
- How often do the forecasts perform better than the reference forecast?

- Are the forecasts better at capturing the interannual variability than the reference forecast?
- Are the forecasts better at discriminating between events and non-events than the reference forecast?
- Are the forecasts sharper than the reference forecast?
- Are the forecasts more sensitive to uncertainty that the reference forecasts?

The verification measures used to answer these questions are described below and summarised in Table 1.

Mean absolute error skill score (MAESS)

One of the most commonly published scores, even the recommended method, when evaluating HEPS is the continuous rank probability score (CRPS, Hersbach, 2000). However, since we have a limited number of data points, only 35 cross validated hindcasts per subbasin, and that the CRPS compares distributions we deemed its use unsuitable for this work. We chose to use the mean absolute error (MAE) to evaluate general forecast performance as the CRPS collapses to the MAE for deterministic forecasts (Hersbach, 2000). Therefore by assuming the ensemble mean to be the deterministic forecast a MAE skill score (MAESS) can be expressed as

$$MAESS = 1 - \frac{MAE_f}{MAE_r} \quad (1)$$

where *f* and *r* denote forecast and reference respectively, and *MAE* is defined as

$$MAE = \frac{1}{n}\sum_{y=1}^{n}\left|\frac{SFV_o^y - SFV_f^y}{SFV_o^y}\right| \quad (2)$$

where *y* denotes year, *n* denotes the total number of years, and *o* denotes observations. The MAESS has a range between negative infinity and 1 with positive values indicating skill over the reference forecast and a value of one a perfect forecast.

Frequency of years (FY^+)

In their work Olsson et al. (2016) proposed FY^+ as a complimentary performance measure to scores such as the MAESS. They are complimentary in that the MAESS is a measure of how much better the forecast is than the reference forecast while FY^+ is the frequency or how often the forecast is better i.e. how often the absolute error is lower. FY^+ scores range from 0 to 100% where values above 50% indicate that the multi-model forecast has skill over the reference forecast. By assuming the ensemble mean to be the deterministic forecast FY^+ is expressed as

$$FY^+ = \frac{100}{n}\sum_{y=1}^{n} H^y$$

where *H* is the Heaviside function defined by

$$H^y = \begin{cases} 0, \ \mathrm{AbsE_r^y} < \mathrm{AbsE_f^y} \\ 1, \ \mathrm{AbsE_r^y} > \mathrm{AbsE_f^y} \end{cases}$$

where *AE* is the absolute error.

Nash–Sutcliffe model efficiency (NSE)

The NSE (Nash and Sutcliffe 1970) is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance. The NSE has a range from -∞ to 1 with 1 being a perfect match and values above 0 denoting that the forecast has skill over climatology. For this work it can be interpreted as how well the forecasted SFV matches the observed SFV year on year and as such is complimentary to MAESS and FY^{+}. By assuming the ensemble mean to be the deterministic forecast the NSE can be expressed as

$$NSE = 1 - \frac{\sum_{y=1}^{n}\left(SFV_o^y - SFV_f^y\right)^2}{\sum_{y=1}^{n}\left(SFV_o^y - \overline{SFV}_o\right)^2} \quad (5)$$

To assess the skill of the multi-model ensemble, with respect to the reference historical ensemble, the difference in their NSE is calculated

$$\Delta NSE = NSE_f - NSE_r \quad (6)$$

where ΔNSE > 0 indicates that the multi-model forecast has skill over the reference forecast.

Relative operating characteristic skill score (ROCSS)

The ROCSS is a skill score based on the area under the curve (AUC) in a relative operating characteristic diagram. ROCSS values below 0 indicate the forecast has no skill over climatology while values over 0 indicate skill with 1 being a perfect forecast. The ROC diagram measures the ability of the forecast ensemble to discriminate between an event and a non-event given a specific threshold. For this work the ROCSS were calculated for the upper tercile ($x \geq 66.7\%$), middle tercile ($66.7\% < x \leq 33.3\%$) and lower tercile ($x < 33.3\%$). These scores estimate the skill of ensemble forecasts to distinguish between below normal (BN), near normal (NN) and above normal (AN) anomalies. Hamill and Juras (2006) define the ROC skill score to be

$$\mathrm{ROCSS} = 2 * \mathrm{AUC} - 1 \quad (7)$$

where AUC is the area under the curve when mapping hit rates against false alarm rates

$$\mathrm{AUC} = \sum_{y=1}^{n+1} \frac{\left(\mathrm{FAR_y} - \mathrm{FAR_{y-1}}\right)\left(\mathrm{HR_y} + \mathrm{HR_{y-1}}\right)}{2} \quad (8)$$

where FAR = false alarm rate and HR = hit rate. False alarms are defined as both the false positive and false negative forecasts, or type I and type II errors. Hits are defined as correctly forecasted events.

Inter quartile range skill score (IQRSS) and uncertainty sensitivity skill score (USS)

Sharpness is an intrinsic attribute to HEPS, giving an indication of how large the ensemble spread is. Forecasts ensembles that are too spread are overly cautious and have limited value for an end user due to the uncertainty of the true magnitude of

the SFV, conversely ensembles that are not spread enough are overly confident and may not be a true representation of the uncertainty thus giving the end user false confidence in the forecast (Gneiting et al., 2007). For this work the sharpness is computed as the difference between the 75th and 25th percentiles of the forecast distribution or the inter quartile range (IQR). The IQRSS is skill score based on the IQR and is a measure of how much better i.e. sharper the forecast ensemble is over the reference ensemble, values above 0 indicate that the forecast ensemble is an improvement over the reference ensemble. The IQRSS is expressed as

$$IQRSS = 1 - \frac{IQR_f}{IQR_r} \tag{9}$$

As mentioned above, sharpness can be misleading. A well-designed and calibrated ensemble should give the user an idea of the uncertainty of the forecast conveyed through the relative sharpness of the ensemble. Thus it follows that the IQR should be positively correlated to the absolute forecast error; a larger (smaller) IQR would indicate to the user that there is a larger (smaller) uncertainty in the SFV forecast. The uncertainty sensitivity skill score (USS) can be expressed as the skill score of the Spearman rank correlations between the IQR and the absolute deterministic error

$$USS = \frac{(\rho_r - \rho_f)}{(1 - \rho_r)} \tag{10}$$

where ρ is the Spearman rank correlation.

2.5 Uncertainty estimation

Due to the limited sample size of data available in this work a bootstrap approach is employed to estimate the verification measures and determine whether they are statistically significant. Again due to data limitations a more circumspect significance level is prudent due to the course nature of the resulting statistics, we chose to set the significance level at 0.1 resulting in a 90% confidence interval between the 5th and 95th percentiles. The cross-validated hindcast ensembles were sampled, allowing for repetition, 10000 times to calculate the verification measures. We define a result to be statistically significant if the 5th (95th) percentile of the bootstrapped ensemble being evaluated does not overlap the 95th (5th) percentile of the bootstrapped reference ensemble.

2.6 Study area and local data

The subbasins used in this work are divided into three groups using the clusters defined by Foster et al. 2012 and Foster et al. 2016, namely clusters, S^1, S^2, and S^3 (Figure 3). Sweden was divided into five regions of homogeneous streamflow variability; three clusters located in the northern parts of the country, where snow dominates the hydrological processes (northern group), and two located in the southern part, where rain dominates the hydrological processes (southern group). For the purposes of this work we are interested only in the northern group. The numbers of subbasins per each of these clusters are 25, 19, and 40 respectively. The S in the cluster's designation denotes that the hydrological regimes are dominated by snow processes and the superscripts give the relative strength of the signal from these processes in the

hydrological regime. During the winter months most of the precipitation that falls within these basins is stored in the form of a snowpack and does not immediately contribute to streamflow. During the warmer spring months, when the temperatures rise above freezing, these snowpacks begin to melt, typically around mid- to late-April, which results in a period of high streamflow commonly referred to as the spring flood. We focus on forecasts of the accumulated streamflow volume during this period or SFV.

For this work, 84 subbasins from seven hydropower producing rivers in northern Sweden (Figure 3) were used to for the development and testing of the multi-model prototype. These are those used in the current operational seasonal forecast system at SMHI plus the two unregulated subbasins used by Olsson et al. (2016). Daily reservoir inflows for each subbasin are available from the SMHI archives starting from 1961 to the end of the last hydrological year; the data used in our work are for the period 1981-2015 due to some of the other datasets used in this work only being available from 1981. These inflows are derived by adding the local streamflow to the change in reservoir storage then subtracting the streamflow from upstream basins i.e.

$$Q_{in} = \Delta S + Q_{local} - Q_{upstream} \quad (11)$$

Missing inflow data were filled by a multiple linear regression approach using simulated inflows for the subbasin and observations from the surrounding subbasins as predictors. Of the 84 subbasins used in this work 68 had less than 1% missing data (50 of these had no missing data), four had 1-10% missing data, five had 20-30% missing data, four had 30-40% missing data, and three had 61%, 63% and 71% missing data respectively. As these subbasins are a part of the current operational forecast system they were included in the study despite some of them having a significant missing data fraction. The average NSE for the data used for filling was 0.70 (the NSE scores for the intervals above were 0.67, 0.75, 0.77, 0.61, and 0.73 respectively) which suggests that this approach is acceptable.

Daily observations of precipitation and temperature data used in this work were obtained from the PTHBV dataset from SMHI (Johansson, 2002). The PTHBV dataset is a 4x4km gridded observation dataset of daily precipitation and temperature data that has been created by optimal interpolation with elevation and wind taken into account. These data are available from 1961 to the present.

Table 2 gives a summary of some basic basin characteristics and statistics regarding the SFV as well as selected performance measures for the HBV rainfall-runoff model for the subbasins in each cluster. Although, the ranges in subbasin areas in the different clusters are similar, except for the maximum in S^3, the SFV statistics increase with each cluster when looking from cluster S^1 to S^3. This is due to the effects that elevation and latitude have on how much snow processes dominate the hydrological regimes in each cluster. Subbasins in cluster S^1 are typically at a latitude or elevation lower than those in clusters S^2 and S^3, similarly the subbasins in S^2 with respect to those in S^3. The ranges in the NSE and the relative MAE imply that in general the HBV model is adequately or well calibrated for most subbasins, however there are some subbasins for which the HBV model appears to not be well calibrated and for which there is some scope for improvement. This can be somewhat misleading as these data are a function of three different observations and as such can be subject to noise and uncertainties.

2.7 Driving Data

Teleconnection indices

This work uses monthly indices of the Arctic Oscillation and Scandinavian Pattern collected from the Climate Prediction Center (Climate Prediction Center, 2012) for the period October 1960 to May 2015.

Seasonal data

The ECMWF seasonal forecast system model from system 4 (Molteni et al., 2011) is the cycle36r4 version of ECMWF IFS (Integrated Forecast System) coupled with a 1° version of the NEMO ocean model. The seasonal forecasts from the ECMWF IFS were used in the following two different forms, a field of seasonal monthly averages as input to the statistical model and individual grid points of daily data for input into the HBV model. The choice to use ECMWF data is primarily a practical one. The ECMWF is an established and proven producer of medium range forecasts and SMHI already has operational access to their products.

The seasonal forecast averages are the seasonal means for each ensemble member of the different predictors which had a domain covering 75°W to 75°E and 80°N to 30°N (Figure 2) with a 1°x1° resolution. For each predictor only the first 15 ensemble members were used in this work. This is because the number of ensemble members available in the ECMWF seasonal forecast is limited to 15 for the hindcast period while the operational seasonal forecast ensemble has 51 members. The predictors considered in this part of the work were the following: 850 hPa geopotential, 850 hPa temperature, 850 hPa zonal wind component, 850 hPa meridional wind component, 850 hPa specific humidity, surface sensible heat flux, surface latent heat flux, mean sea level pressure, 10m zonal wind component, 10m meridional wind component, 2m temperature, total precipitation.

The daily time series data are the ECMWF IFS seasonal forecasts of daily values of temperature (2mT) and the accumulated total precipitation (pr). These data have a resolution of 0.5°x0.5° and spans a period from 1981 to 2015 and had a domain covering 11° to 23°E and 55° to 70°N.

3 Results and Discussion

The following section outlines and discusses the results from the cross-valuated hindcasts of the different approach's ability to hindcast the SFV for the period 1981-2015. First this evaluation is done for each system using climatology as a reference to assess their general skill. After that the more skilful of the two multi-model ensembles is evaluated using the HE as a reference to assess any improved skill and thus added value of the multi-model ensemble approach over the current HE approach. The analysis is carried out on the cross-validated hindcasts of the SFV initialised on the 1st of January, February, March, April, and May.

3.1 Evaluating the different forecast systems against climatology

The different forecast approach's general skill to predict the SFV was estimated using MAESS, their skill to reproduce the interannual variability was estimated using NSE, and finally the skill to discriminate between below BN, NN, and AN SFVs is estimated using ROCSS. Table 3 gives an overview of these scores across all subbasins and clusters for each initialisation month as well as the percentage of subbasins where the hindcasts outperformed climatology, the values in brackets are the percentage of subbasins where the hindcasts outperformed climatology and the result is statistically significant.

The performance measures for each of the three approaches are positively related to the relative timing of the hindcasts i.e. hindcasts initialised in any month are generally more (less) skilful than the hindcasts initialised in the preceding (following) months. This can be expected as the further away in time from the spring flood that the hindcast is initialised, increasing lead time, the less the initial hydrological conditions contribute to predictability and the more uncertain the forcing data become (e.g. Wood et al., 2016; Arnal et al., 2017).

With respect to general skill and the ability to capture the interannual variation shown by the observations, the prototypes tend to perform better than HE with ME_{hds} typically having the best performance. This is especially so when we consider the percentage of the subbasins where this improved performance is statistically significant. The gap between HE and the two prototypes in MAESS, NSE, and percentage of subbasins with improved performance over climatology tends to get smaller as the season progresses while the gap in the percentage of subbasins where improved performance is statistically significant appears to grow, at least early in the season.

However, if we turn our attention to the forecast's ability to discriminate between BN, NN, and AN SFVs then the HE holds an advantage over the two prototypes especially when it comes to identifying NN events from all forecast initialisation dates and, to a lesser extent, BN events for the later forecasts. The proposed prototypes are better at identifying AN events for all forecasts except those initialised in May where the ability of the HE is comparable. The advantage displayed by the HE to identify NN events is to be expected due to its climatological nature while the advantage with respect to BN events can probably be attributed to a cold bias in the historical forcing data caused by climate change. The drop in relative skill by the prototypes in the later forecasts is in part due to their sharpness being worse than the HE in the later forecast (Section 3.3).

From these results we are now able to make an informed choice as to which prototype to proceed with, ME_{hds} (hereafter referred to as the prototype unless stated otherwise). If we take all the results and rank the performances of the three methods then the prototype would rank first followed closely by ME_{ads} and HE would rank third. However, all three forecast methods have been shown to be skilful at forecasting the SFV albeit a naïve skill.

3.2 Evaluating the prototype against HE

The frequency at which the prototype outperforms HE is estimated using FY^{+}, its general skill to predict the SFV is estimated using MAESS, its skill to reproduce the interannual variability is estimated using ΔNSE, and finally its skill to discriminate between BN, NN, and AN SFVs is estimated using ΔROCSS. Figure 4 shows the bootstrapped scores for

MAESS, FY+, and ΔNSE calculated for each hindcast initialisation month for the subbasins in cluster S^3. The medians of these bootstrapped scores are presented as a histogram; summary statistics are documented above the histogram. On the left-hand side are the max, mean, and min scores for the cluster i.e. the subbasins with the highest and lowest scores and the mean for the basin. On the right-hand side are the percentage of subbasins where the prototype outperformed HE, shows skill over HE (n^{+}_{abs}), the percentage of subbasins where the prototype shows statistically significant skill over HE ($n^{+}_{0.1}$), and the percentage of subbasins where HE statistically significant skill over the prototype ($n^{-}_{0.1}$).

FY^+

The prototype has a $FY^+ > 50\%$ for the majority of the subbasins in cluster S^3, ranging from 98% of the subbasins with a mean FY^+ of 61% for hindcasts initialised in January down to 73% of the subbasins and mean FY^+ of 56% for hindcasts initialised in May. These figures are similar, even a little higher, for subbasins in cluster S^2 while somewhat lower for cluster S^1. The number of subbasins for which the prototype has a statistically significant $FY^+ > 50\%$ ranges between 10% and 28% in cluster S^3 (S^2 = 5-37%, and S^1 = 12-16%). While the prototype has a statistically significant $FY^+ < 50\%$ (performs worse than HE more often than not) for 5% of subbasins for hindcasts initialised in April in cluster S^3 and 4% of subbasins in hindcasts initialised in May for cluster S^1.

MAESS

The prototype shows skill at improving the volume error hindcasted by HE for the majority of the subbasins, ranging between 65% and 100% of the subbasins in cluster S^3 (S^2 = 74-95%, and S^1 = 64-80%). This improvement tends to be largest for hindcasts initialised in January, mean MAESS of 0.12 (S^2 = 0.11, and S^1 = 0.04), and lowest for those in May, 0.04 (S^2 = 0.05, and S^1 = 0.02). The percentage of subbasins for which MAESS > 0 is statistically significant ranges between 10-53% for all clusters and hindcast initialisations, while the percentage of subbasins for which MAESS < 0 is statistically significant are 8% and 4% for hindcasts initialised in March and May in cluster S^1 and 3% for hindcasts initialised in both April and May in cluster S^3. These results also show that the prototype generally has a smaller MAE than HE especially for earlier hindcast initialisations and again for clusters S^3 and S^2.

ΔNSE

The prototype shows skill at improving the representation of the interannual variability of the observed SFV again for most of the subbasins, ranging between 63% and 100% of subbasins in cluster S^3 (S^2 = 74-100%, and S^1 = 76-92%), and the mean ΔNSE ranges between 0.06 and 0.33 for subbasins in cluster S^3 (S^2 = 0.09 and 0.32, and S^1 = 0.06 and 0.15). The percentage of subbasins for which ΔNSE > 0 is statistically significant ranges between 16% and 63% for all clusters and hindcast initialisations, while the percentage of subbasins for which ΔNSE < 0 is statistically significant are 4% for hindcasts initialised in January, March and May in cluster S^1, and 5% for hindcasts initialised in May in cluster S^3.

ΔROCSS

Figure 5, which has the same information presentation structure as Figure 4, shows the bootstrapped ΔROCSS for the lower (BN), middle (NN), and upper (AN) terciles calculated for each hindcast initialisation month for the subbasins in cluster S^3.

The prototype shows skill over HE to discriminate between BN events and non-BN events for the majority of the subbasins in cluster S^3 for hindcasts initialised in January and February, 95% and 68% respectively (S^2 = 63%, 53% and S^1 = 68%, 32%) but this drops to less than half the subbasins in hindcasts initialised thereafter. The mean ΔROCSS ranges between -0.04 and 0.14 in cluster S^3 (S^2 = -0.03 and 0.05, and S^1 = -0.06 and 0.02) with only statistically significant results being found in favour of the prototype for 15% and 5% of subbasins for hindcasts initialised in January in clusters S3 and S2 respectively and in favour of HE for 4% of subbasins for hindcasts initialised in April in cluster S^1.

Out of the three terciles the prototype shows the least skill over HE at discriminating between NN events and non-NN events. The percentage of subbasins for which the prototype outperforms HE ranges between 23% and 57% (S^2 = 37-63%, and S^1 = 24-60%) with mean ΔROCSS ranges between -0.07 and 0.01 (S^2 = -0.06 and 0, and S^1 = -0.05 and 0.02) and no statistically significant results for any subbasins, both in favour or against the prototype.

The prototype shows the best performance when discriminating between AN events and non-AN events. The percentage of subbasins for which the prototype shows skill over HE in the upper tercile ranges between 85% and 98% for hindcasts initialised in the first three months (S^2 = 47-89%, and S^1 = 48-84%) then 57% and 25% for the last two months respectively (S^2 = 63% and 53%, and S^1 = 80% and 32%). The mean ΔROCSS ranges between -0.02 and 0.13 (S^2 = -0.01 and 0.14, and S^1 = -0.01 and 0.07). The percentage of subbasins in cluster S^3 for which ΔROCSS > 0 is statistically significant are 18%, 10%, and 3% for hindcasts initialised in January, February, and March respectively, and 16% for forecasts initialised in January and February in cluster S^2. There are no statistically significant results in favour of HE.

3.3 Analysis of the forecast ensemble sharpness

Figure 6 shows the cross validated hindcasts by the prototype initialised in January (top panel) and May (bottom panel) for Göuta-Ajaure, a cluster S^3 subbasin in upper reaches of the Ume River system. This basin was chosen as an example of a where the prototype showed typical performance results i.e. neither the best nor the worst. The total ensemble spread (the whiskers) of the forecasts initialised in January remains somewhat consistent from year to year while the IQR (the blue boxes) displays a more pronounced variation. The lack of variation in total spread is primarily the result of the climatological nature of the HE component which tends to have a larger and more consistent spread than that for DE and SE at longer lead times. The greater variation exhibited by the IQR is mostly due to the 'true' forecast nature of the DE and SE components in the multi-model ensemble. If we turn our attention to the forecasts initialised in May we see a more pronounced variation in both the total spread and the IQR. This is because the spread in the DE and SE components is now comparable to and often larger than the spread in the HE component. Table 4 shows how the IQRSS drops as the spring flood season approaches. It can also be seen in figure 6 that the ensemble median (red line) is more responsive to the year on year variation in SFV in the May forecasts than in the January forecasts. This is because the relative contribution to predictability by the initial conditions is greater than the contribution from the meteorological drivers closer to the spring flood period. These patterns are generally true for both the forecasts initialised in the intermediate months and for the other subbasins.

If we assume that the more sensitive an ensemble is to uncertainty the more the forecast sharpness will vary. We would therefore expect the USS values to generally be positive i.e. that the forecast sharpness of prototype is better correlated with the forecast error than for HE. This is largely supported by the USS values in Table 4 where only three values are negative, the January forecast in cluster S^2 and the April forecasts in both clusters S^1 and S^2, and even then not by very much. This suggests that at least one but probably both of the DE and SE ensembles are responsible for this improvement due to their variability. There is a general decreasing trend with initialisation date in the USS values in clusters S^1 and S^2 (if we ignore the value for January in S^2) while the values are more consistent in cluster S^3. All the uncertainty correlation values for both the HE and the prototype are significant at the 0.1 level (not shown for brevity) suggesting that both exhibit sensitivity to uncertainty to some extent, however the prototype is generally more so which should instil more confidence for the forecast in the users.

The IQRSS values show that the prototype tends to produce sharper forecasts than HE early in the season i.e. for forecasts initialised in January and February in cluster S1 and forecasts initialised in January, February and March in clusters S2 and S3. This is reversed for the remaining initialisation dates where HE tends to produce sharper forecasts than the prototype.This is probably due to the climatological nature of the HE having less of an impact on forecast sharpness as the initialisation date approaches the spring flood period together with the uncertainties and biases in the other individual ensembles exacerbating the situation.

3.4 Spatial and temporal variations and transferability of the prototype

Both multi-model ensembles show skill at forecasting SFV with respect to forecast error, ability to reproduce the interannual variability in SFV, and the ability to discriminate between BN, NN, and AN events. The prototype, in particular, is at worst comparable to the HE and at best clearly more skilful. This relative performance of the prototype varies both in space and time. Figure 7 shows maps of the median bootstrapped FY^+ values. For hindcasts initialised in January the spatial pattern in the FY+ scores show that the prototype tends to outperform HE more in subbasins that have a higher latitude or elevation. However, as the initialisation date approaches the spring flood period this pattern becomes less and less coherent. This general pattern is also true for MAESS scores. This suggests that the change in the performances of the prototype and HE, as a function of initialisation date, are not always similar for subbasins that are near one another. Further work would be needed to find out what the underlying reason for this is.

Data availability is the biggest limiting factor to the transferability of this approach to other areas. The HE, AE, and SE approaches are all dependant on good quality observation time-series. Additionally, the skill all three of these approaches would be expected to be affected by length of these time-series. They length of the time-series should be long enough to be a good representative sample of the climatology otherwise the forecasts would be biased in favour of the climate represented in the data and not the true climatology.

The SE and AE approaches require an understanding of how the variability in the local hydrology is affected by large scale circulation phenomena such as teleconnection patterns to help select predictors and teleconnection indices for inputs to each

approach respectively. The hydrological rainfall-runoff model used in the prototype should not pose a problem, although HBV has been successfully setup for snow dominated catchments outside of Sweden (e.g. Seibert et al., 2010; Okkonen and Kløve, 2011), any sufficiently well calibrated rainfall-runoff model would suffice.

We believe that, if the above requirements are met, a seasonal hydrological forecast system similar to the prototype can be setup in other snow dominated regions around the world.

4 Conclusions

In this paper we present the development and evaluation of a hydrological seasonal forecast system prototype for predicting the SFV in Swedish rivers. Initially, two versions of the prototype, ME_{ads} and ME_{hds}, were evaluated together with the HE using climatology as a reference to both help select which version of the prototype to proceed with and to get a general impression of their skill to forecast the SFV. Thereafter the chosen prototype was evaluated using HE as a reference and finally the sharpness of the hindcast ensembles were analysed.

The main findings are summarized below:

- The prototype is able to outperform the HE approach 57% of the time on average. It is at worst comparable to the HE in forecast skill and at best clearly more skilful.
- The prototype is able to reduce the forecast error by 6% on average. This translates to an average volume of 9.5 x 10^6 m^3.
- The prototype is generally more sensitive to uncertainty, that is to say that the ensemble spread tends to be more correlated with the forecast error. This is potentially useful to users as the ensemble spread could be used as a measure of the forecast quality.
- The prototype is able to improve the prediction of above and below normal events early in the season.

Looking forward, future studies need to address the questions raised by Zhao et al. (2017) regarding the bias adjustment of meteorological seasonal forecast data using quantile mapping. Results from this study show that while the seasonal forecasts were bias adjusted the performance of the DE was disappointing, although it still had value within the multi-model setting suggesting that it has more of a modulating roll on the other modelling chains as opposed to contributing directly to predictability.

How the individual model ensembles are combined to give the multi-model output needs to be revisited. When we applied the asymmetric weighting scheme proposed by Olsson et al. (2016) we did not find that it improved the multi-model performance in general across all stations and forecasts and so did not use it. However, we do believe that more work should be done to find a more appropriate weighting scheme than simple pooling. Perhaps by better understanding how the

performance of the different modelling chains are affected by the initial conditions and lead-time it will shed more light on how to best approach this issue. Further development and testing along these lines are planned for the future.

The AE approach did not exhibit the promising performances found by Olsson et al. (2016) using circulation pattern analysis to select the analogues. A part of the explanation for this poor performance is that the teleconnection information used to select the analogues only partially span the full periods Foster et al. (2016) identified, from October/November to the beginning of the spring flood. The missing data could be filled by making forecasts of the indices. Another approach would be to revisit the circulation pattern analysis based approach now that data inhomogeneity issues are largely addressed by the new ERA5 reanalysis data that is becoming available (http://climate.copernicus.eu/products/climate-reanalysis). Yet another approach would be to use GCM forecasts to select the analogues (e.g. Crochemore et al., 2016).

Lastly, the post processing of model outputs (e.g. Lucatero et al., 2017) has been shown to be beneficial, the incorporation of a simple approach like linear scaling is possibly the most appealing due to its ease of implementation in an operational environment.

Data availability. ECMWF seasonal forecasts are available under a range of licences, for more information visit http://www.ecmwf.int. AO and SCAND teleconnection indices are available for download from the Climate Prediction Center website (http://www.cpc.ncep.noaa.gov). For streamflow data and the PTHBV dataset please contact customer services at SMHI (customerservice@smhi.se).

Author contributions. The AE and SE approaches were designed by K. Foster and C.B. Uvo and were implemented by K. Foster. The DE approach was designed by J. Olsson and implemented by K. Foster. The multi-model experimental set-up was designed and implemented by K. Foster. The manuscript was prepared by K. Foster with contributions from C.B. Uvo and J. Olsson.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by research projects funded by Energiforsk AB (formerly Elforsk AB) and the EUPORIAS (EUropean Provision Of Regional Impacts Assessments on Seasonal and decadal timescales) project funded by the Seventh Framework Programme for Research and Technological Development (FP7) of the European Union (grant agreement no. 308291). Many thanks go to Johan Södling, Jonas German, and Barbro Johansson for technical assistance as well as fruitful discussions. Constructive and detailed reviews of the original manuscript by Joost Beckers, Fernando Mainardi Fan, Hannes Tobler, and Sebastian Röthlin are gratefully acknowledged.

References

Arheimer, B., Lindström, G., & Olsson, J.: A systematic review of sensitivities in the Swedish flood-forecasting system. Atmospheric Research, 100(2), 275-284, 2011.

Arnal, L. , Wood, A. W., Stephens, E., Cloke, H. L. and Pappenberger, F.: An efficient approach for estimating streamflow forecast skill elasticity. J. Hydrometeorol., 18, 1715-1729, 2017.

Beckers, J. V. L., Weerts, A. H., Tijdeman, E., and Welles, E.: ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction, Hydrol. Earth Syst. Sci., 20, 3277-3287, https://doi.org/10.5194/hess-20-3277-2016, 2016.

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resources Research, 52, 8238-8259, 2016.

Berg, P., Bosshard, T., and Yang, W.: Model consistent pseudo-observations of precipitation and their use for bias correcting regional climate models. Climate, 3(1), 118-132, 2015.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, SMHI Reports RHO No. 7, SMHI, Norrköping, Sweden, 1976.

Cheng, X., & Dunkerton, T. J.: Orthogonal rotation of spatial patterns derived from singular value decomposition analysis, J. Clim., 8, 2631-2643, 1995.

Climate Prediction Center: Teleconnections, Arctic oscillation (AO). http://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao_index.html, last access: 12 November 2016, 2005.

Climate Prediction Center: Teleconnections, Scandinavia (SCAND). http://www.cpc.ncep.noaa.gov/data/teledoc/scand.shtml, last access: 12 November 2016, 2012.

Cloke HL, Pappenberger F.: Evaluating forecasts for extreme events for hydrological applications: an approach for screening unfamiliar performance measures. Meteorological Applications 15(1): 181–197, 2008.

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

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

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R.: Seasonal climate predictability and forecasting: status and prospects. Wiley Interdisciplinary Reviews: Climate Change, 4(4), 245-268, 2013.

ECMWF: IFS Documentation, http://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation, last access: 15 April 2017, 2016.

Foster, K.L., & Uvo, C.B.: Seasonal streamflow forecast: a GCM multimodel downscaling approach, Hydrology Research 41:503-507, 2010.

Foster, K. and Uvo, C. B.: Regionalisation of Swedish hydrology, XXVII Nordic Hydrology Conference, Oulu, 2012.

Foster, K., Uvo, C.B., & Olsson, J.: The spatial and temporal effect of selected teleconnection phenomena on Swedish hydrology. Submitted to Clim. Dyn., 2016.

Garen, D.: Improved techniques in regression-based streamflow volume forecasting, J. Water Resour. Pl. Manage., 118, 654–670, 1992.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. Roy. Stat. Soc. B, 69, 243– 268, doi:10.1111/j.1467-9868.2007.00587.x, 2007.

Hamill TM, Juras J.: Measuring forecast skill: is it real skill or is it the varying climatology? Q. J. R. Meteorol. Soc. 132: 2905–2923, 2006.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15, 559–570, 2000.

Johansson, B.: Estimation of areal precipitation for hydrological modelling, PhD thesis, Earth Sciences Centre, Report no. A76, Göteborg University, Göteborg, 2002.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S.: Development and test of the distributed HBV-96 hydrological model. Journal of hydrology, 201(1), 272-288, 1997.

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J., and Jensen, K. H.: Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: effect of preprocessing and postprocessing on skill and statistical consistency, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-379, 2017.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Tech. Memo., 656, 49 pp., available at: http://www.ecmwf.int/sites/default/files/elibrary/2011/ 11209-new-ecmwf-seasonal-forecast-system-system-4.pdf (last access: 29 August 2016), 2011.

Nash, J. E.; Sutcliffe, J. V,: "River flow forecasting through conceptual models part I — A discussion of principles". Journal of Hydrology. 10 (3): 282–290. doi:10.1016/0022-1694(70)90255-6, 1970.

Nilsson, P., Uvo, C. B., Landman, W. A., & Nguyen, T. D.: Downscaling of GCM forecasts to streamflow over Scandinavia. Hydrology Research, 39(1), 17-26, doi:10.1016/j.jhydrol.2005.08.007, 2008.

Okkonen, J., and Kløve, B.: A sequential modelling approach to assess groundwater–surface water resources in a snow dominated region of Finland. Journal of hydrology, 411(1-2), 91-107, 2011.

Olsson, J., Yang, W., Graham, L.P., Rosberg, J., and Andréasson J.: Using an ensemble of climate projections for simulating recent and near-future hydrological change to Lake Vänern in Sweden, Tellus, 63A, 126-137, doi:10.1111/j.1600-0870.2010.00476.x, 2011.

Olsson, J., Uvo, C. B., Foster, K., & Yang, W.: Technical Note: Initial assessment of a multi-method approach to spring-flood forecasting in Sweden. Hydrology and Earth System Sciences, 20(2), 659-667, 2016.

Pagano, T. C., Garen, D. C., Perkins, T. R., and Pasteris, P. A.: Daily updating of operational statistical seasonal water supply forecasts for the western U.S., J. Am. Water Resour. As., 45, 767–778, 2009.

Schepen, A., Zhao, T., Wang, Q. J., Zhou, S. and Feikema, P.: Optimising seasonal streamflow forecast lead time for operational decision making in Australia, Hydrol. Earth Syst. Sci., 20(10), 4117–4128, doi:10.5194/hess-20-4117-2016, 30 2016.

Seibert, J., McDonnell, J. J., and Woodsmith, R. D.: Effects of wildfire on catchment runoff response: a modelling approach to detect changes in snow-dominated forested catchments. Hydrology research, 41(5), 378-390, 2010.

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

Statistiska centralbyrån: Electricity supply, district heating and supply of natural gas 2015, Report EN11SM1601, Statistiska centralbyrån, Stockholm, Sweden, 2016.

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

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, Journal of Hydrometeorology, 17, 651-668, 2015.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, L14,01, doi:10.1029/2008GL034648, 2008.

Yang, W., Andréasson, J., Graham, L. P., Olsson, J., Rosberg, J., & Wetterhall, F.: Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. Hydrology Research, 41(3-4), 211-229, 2010.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resour. Res., 49, 4687–4699, doi:10.1002/wrcr.20350, 2013.

Yossef , N.C., van Beek, R., Weerts, A., Winsemius, H. and Bierkens, M.F.P. Skill of a global forecasting system in seasonal ensemble streamflow prediction. Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-604, in review, 2016.

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

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

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

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

Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M. -H. Ramos,: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? Journal of Climate, 30, 3185-3196, doi:10.1175/JCLI-D-16-0652.1, 2017.

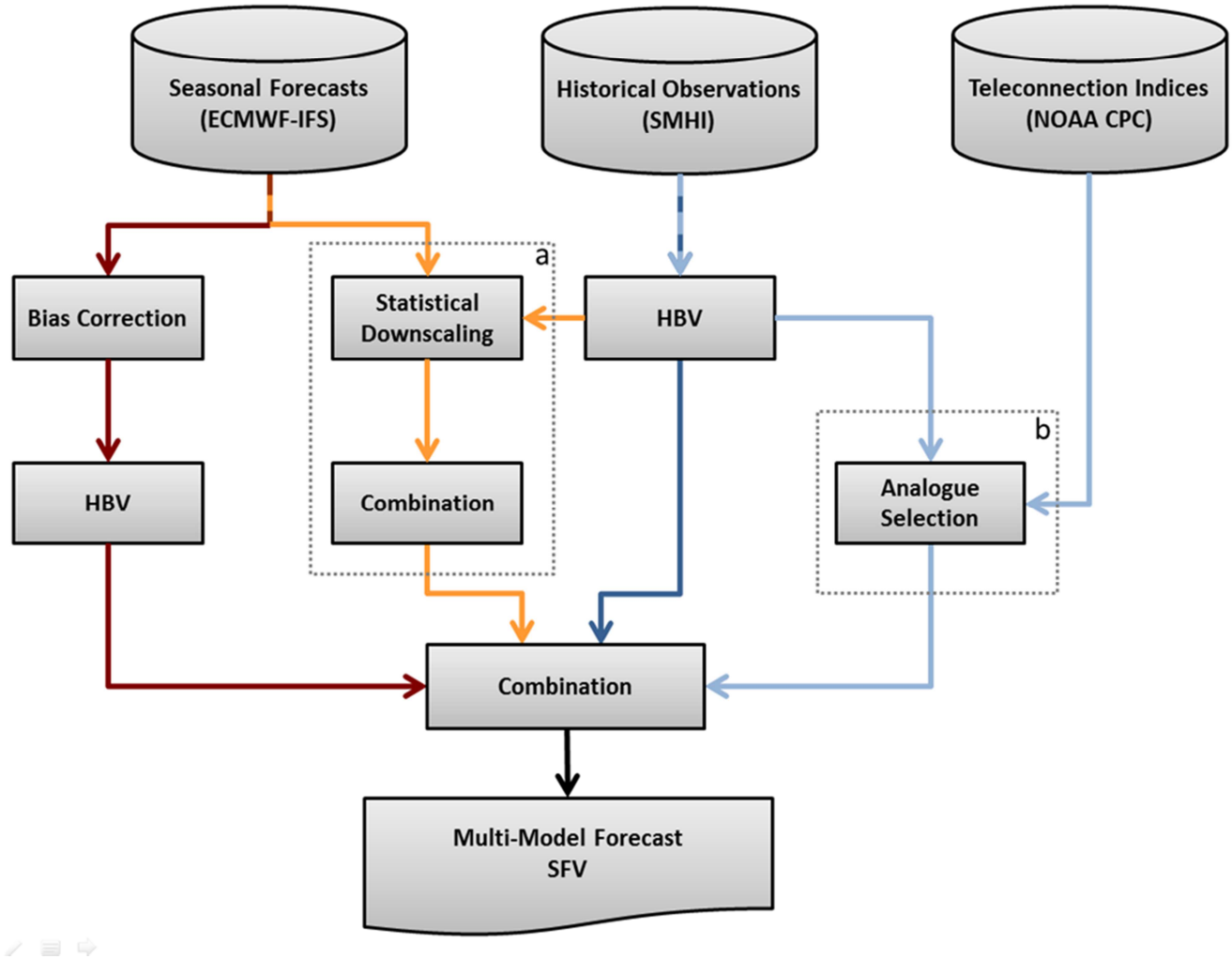

Figure 1. Schematic of the multi-model forecast system. The three individual model chains that are included in the multi-model are (from left to right) the dynamic model chain (red lines), the statistical model chain (orange lines), and the historical (dark blue lines) or analogue (light blue lines) model chain. The dashed boxes labelled (a) and (b) indicate the parts of the system that have non trivial changes from the multi-model described in Olsson et al. (2016).

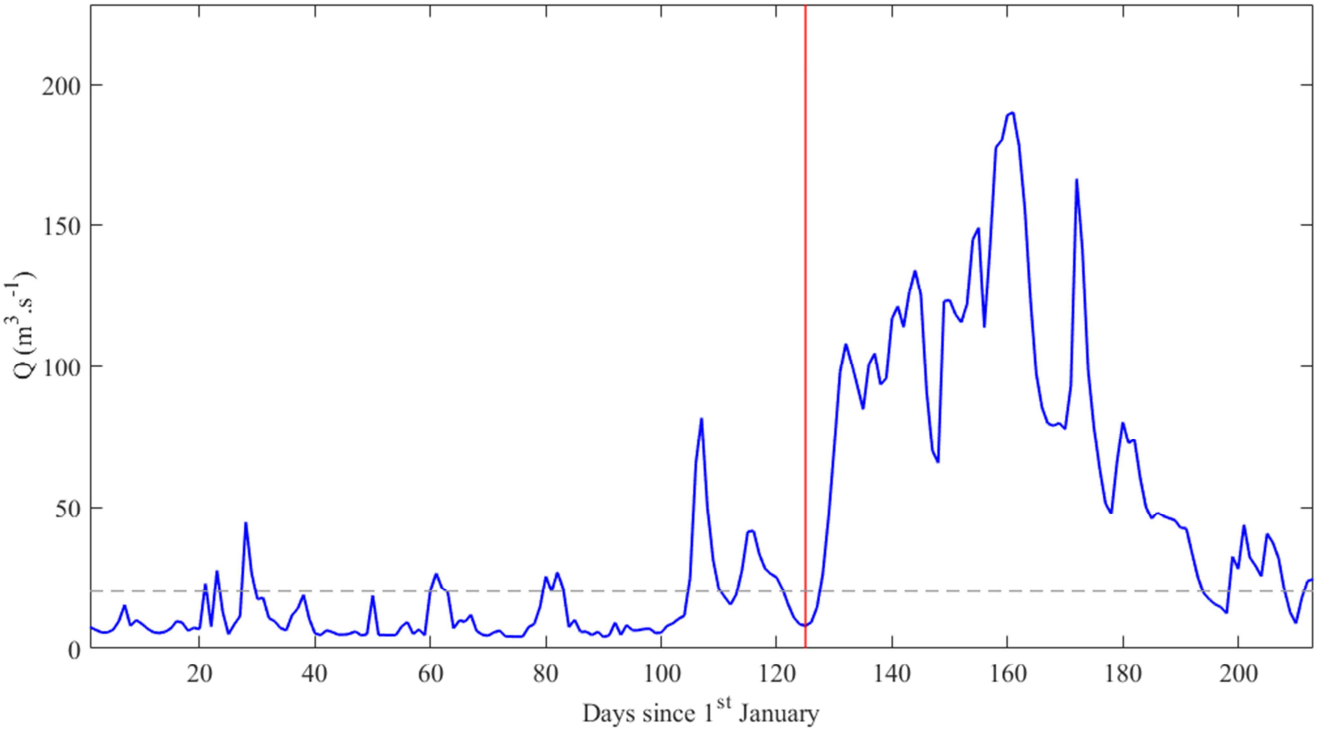

Figure 2. Schematic of how the spring flood is defined. The spring flood is the period between the onset and the last day of July (day 211/212 since the 1st January). The hydrograph from which the spring flood period is to be derived (blue line), the onset date (red line), and the 90th percentile of the inflow for first 80 days (dashed line).

Table 1. The validation metrics used to evaluate the multi-model performance. The threshold for skill is 50 for FY^+ and 0 for all the other metrics.

Name	Equation	Description
Mean absolute error skill score (**MAESS**)	$$\boldsymbol{MAESS = 1 - \frac{MAE_f}{MAE_r}}$$	Measure of the model's general performance; it quantifies the relative forecast error against a reference forecast.
Frequency of Years ($\mathbf{FY^+}$)	$$\boldsymbol{FY^+ = \frac{100}{n}\sum_{y=1}^{n} H^y},$$ where H is the Heaviside function defined by $$H^y = \begin{cases} 0, & AE_r^y < AE_f^y \\ 1, & AE_r^y > AE_f^y \end{cases},$$ AE is the absolute error.	Measure of the model's general performance; it quantifies how often the forecast outperforms a reference forecast.
Nash-Sutcliffe efficiency (**NSE**)	$$\boldsymbol{NSE = 1 - \frac{\sum_{y=1}^{n}\left(SFV_{obs}^y - SFV^y\right)^2}{\sum_{y=1}^{n}\left(SFV_{obs}^y - \overline{SFV}_{obs}\right)^2}}$$	Measure of the model's general performance; it quantifies the model's residual variance against a reference forecast's variance.
Relative operating characteristic skill score (**ROCSS**)	$$\boldsymbol{ROCSS = 2 * AUC - 1},$$ where AUC is the area under the curve $$AUC = \sum_{y=1}^{n+1}\frac{(FR^y - FR^{y-1})(HR^y + HR^{y-1})}{2},$$ where FR is the false alarm rate and HR is the hit rate.	Measure of the model's probabilistic performance; it quantifies the model's ability of the discriminate between an event and a non-event given a specific threshold.
Interquartile range skill score (**IQRSS**)	$$\boldsymbol{IQRSS = 1 - \frac{IQR_f}{IQR_r}}$$ where IQR is the interquartile range.	Measure of the forecast sharpness, it quantifies the relative spread in the forecast against a reference forecast.
Uncertainty sensitivity skill score (**USS**)	$$\boldsymbol{USS = \frac{(\rho_r - \rho_f)}{(1 - \rho_r)}},$$ where ρ is the Spearman rank correlation between the IQR and absolute error.	Measure of the model's sensitivity to uncertainty; it quantifies the correlation between forecast sharpness and absolute error

Table 2. Basic information on the study area including overall performance of the HBV model for the subbasins in each cluster.

Cluster		Basin		SFV			HBV	
		Area	elevation	(m^3 x 10^8)			NSE	rMAE
		(km^2)	(m)	min	mean	max		(%)
1	min	233	135	0.21	0.42	0.82	-0.47	3.9
	median	1827	282	1.27	3.23	5.30	0.74	11.2
	max	6258	584	18.95	34.36	44.36	0.95	44.5
2	min	184	429	0.40	0.81	1.68	-0.69	6.2
	median	1166	598	2.49	3.80	4.77	0.66	10.2
	max	4272	666	9.99	13.56	18.17	0.83	70.1
3	min	270	212	0.67	1.12	1.54	0.22	3.8
	median	1309	586	2.96	5.41	7.95	0.74	7.3
	max	13177	776	19.39	37.76	48.50	0.92	20.0

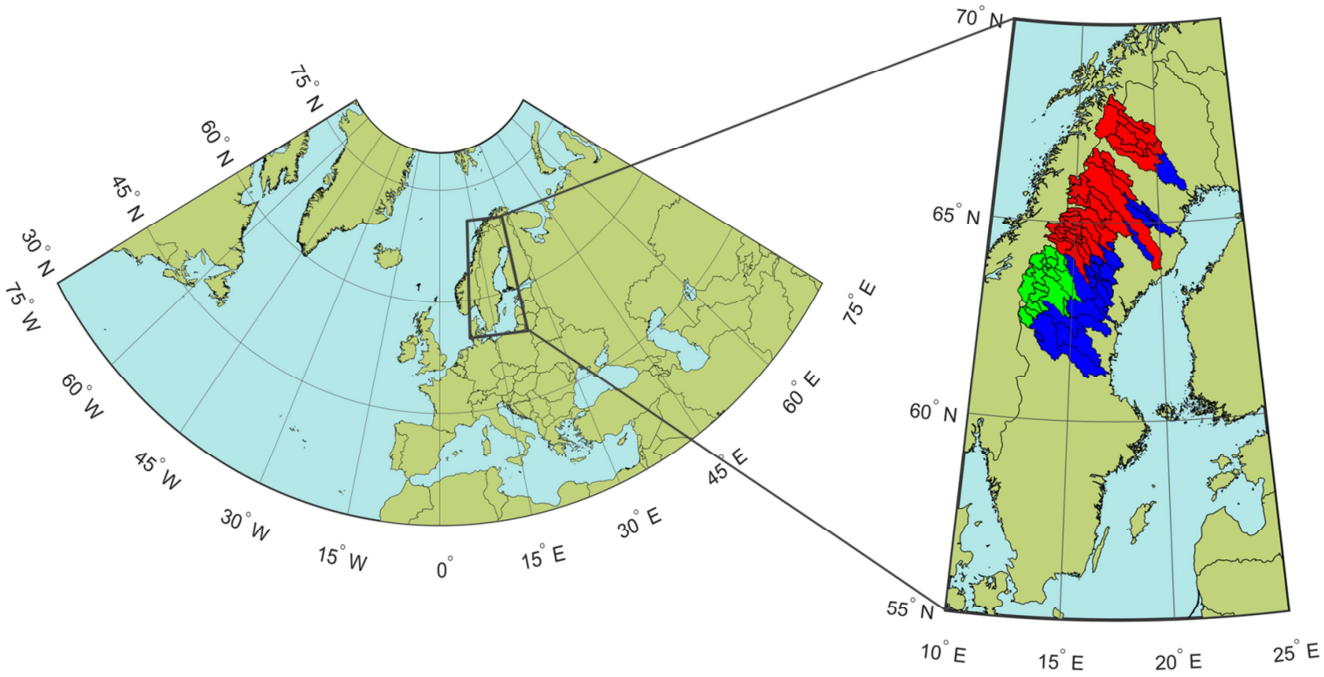

Figure 3. Map showing (left) the domain for the predictors used in the SE modelling chain, (right) the domain of the seasonal forecast data used in the DE modelling chain, and the location of the forecasts subbasins used in this work. The subbasins shown in blue belong to cluster S^1, those shown in green belong to cluster S^2, and those shown in red belong to cluster S^3.

Table 3. Bootstrapped (N = 10000) skill scores and the number of subbasins, as a percentage, where the HEPS performs better than climatology averaged over all 84 subbasins. The n^+ values in brackets show the percentages of the subbasins for which these scores are statistically significant at the 0.1 level.

		MAESS		NSE		ROCSS					
						BN		NN		AN	
			n^+ (%)		n^+ (%)		n^+ (%)		n^+ (%)		n^+ (%)
HE	Jan	-0.09	25 (1)	-0.24	17 (0)	0.23	90 (21)	0.07	70 (0)	0.10	68 (11)
	Feb	0.00	51 (6)	-0.07	42 (5)	0.41	99 (52)	0.11	69 (1)	0.26	92 (27)
	Mar	0.09	80 (17)	0.13	77 (23)	0.55	100 (87)	0.10	73 (5)	0.44	99 (56)
	Apr	0.15	85 (35)	0.22	80 (35)	0.62	100 (92)	0.17	85 (7)	0.51	100 (75)
	May	0.21	90 (49)	0.32	90 (49)	0.68	100 (98)	0.23	92 (10)	0.61	100 (92)
ME_{ads}	Jan	0.00	50 (2)	0.00	55 (1)	0.31	99 (31)	-0.01	48 (0)	0.20	80 (18)
	Feb	0.06	73 (23)	0.11	76 (21)	0.39	99 (51)	0.08	74 (0)	0.36	96 (42)
	Mar	0.11	86 (25)	0.20	87 (36)	0.47	100 (76)	0.07	61 (4)	0.47	100 (60)
	Apr	0.20	95 (62)	0.32	94 (64)	0.60	100 (90)	0.16	83 (5)	0.52	100 (79)
	May	0.22	96 (67)	0.36	98 (68)	0.66	100 (94)	0.18	82 (8)	0.57	100 (76)
ME_{hds}	Jan	0.02	60 (6)	0.03	63 (5)	0.32	100 (31)	0.00	51 (0)	0.22	83 (24)
	Feb	0.08	80 (25)	0.14	85 (29)	0.41	99 (57)	0.07	69 (1)	0.38	99 (44)
	Mar	0.14	90 (32)	0.24	92 (45)	0.51	100 (81)	0.07	61 (5)	0.48	100 (64)
	Apr	0.19	94 (56)	0.32	93 (62)	0.60	100 (90)	0.17	88 (5)	0.54	100 (80)
	May	0.24	98 (74)	0.39	96 (76)	0.67	100 (94)	0.18	85 (10)	0.60	100 (88)

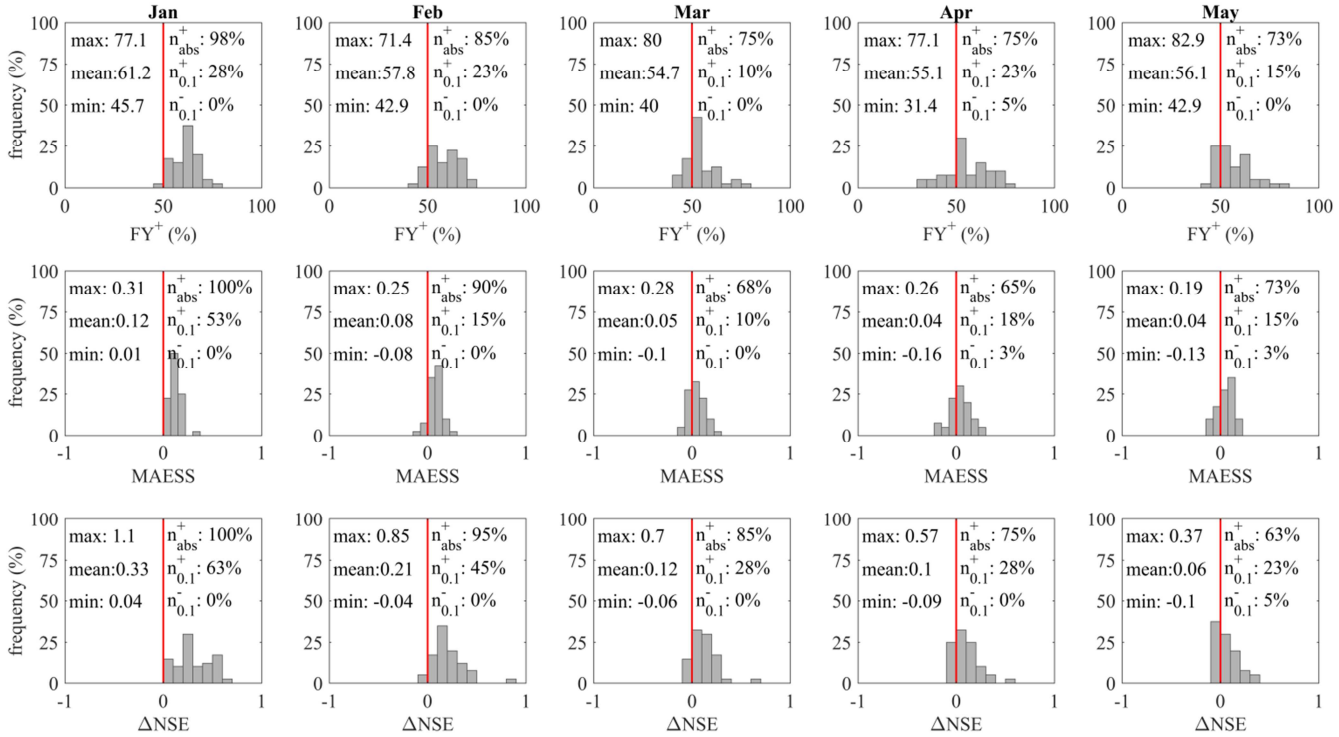

Figure 4. Bootstrapped (N = 10000) FY+, MAESS, and ΔNSE scores for ME_{hds} with respect to HE for all subbasins in the cluster S^3. Each subplot is a histogram of the medians of the bootstrapped validations scores for each initialisation month. Above the histograms are six related statistics: (left of the red line) the maximum, mean, and minimum of the validation scores shown in the histograms; (right of the red line) percentages of the subbasins where ME_{hds} performed better than HE (n^+_{abs}), the percentage of subbasins where ME_{hds} performed better than HE ($n^+_{0.1}$) at the significance level 0.1, and lastly the percentage of subbasins where ME_{hds} performed worse than HE ($n^-_{0.1}$) at the 0.1 level.

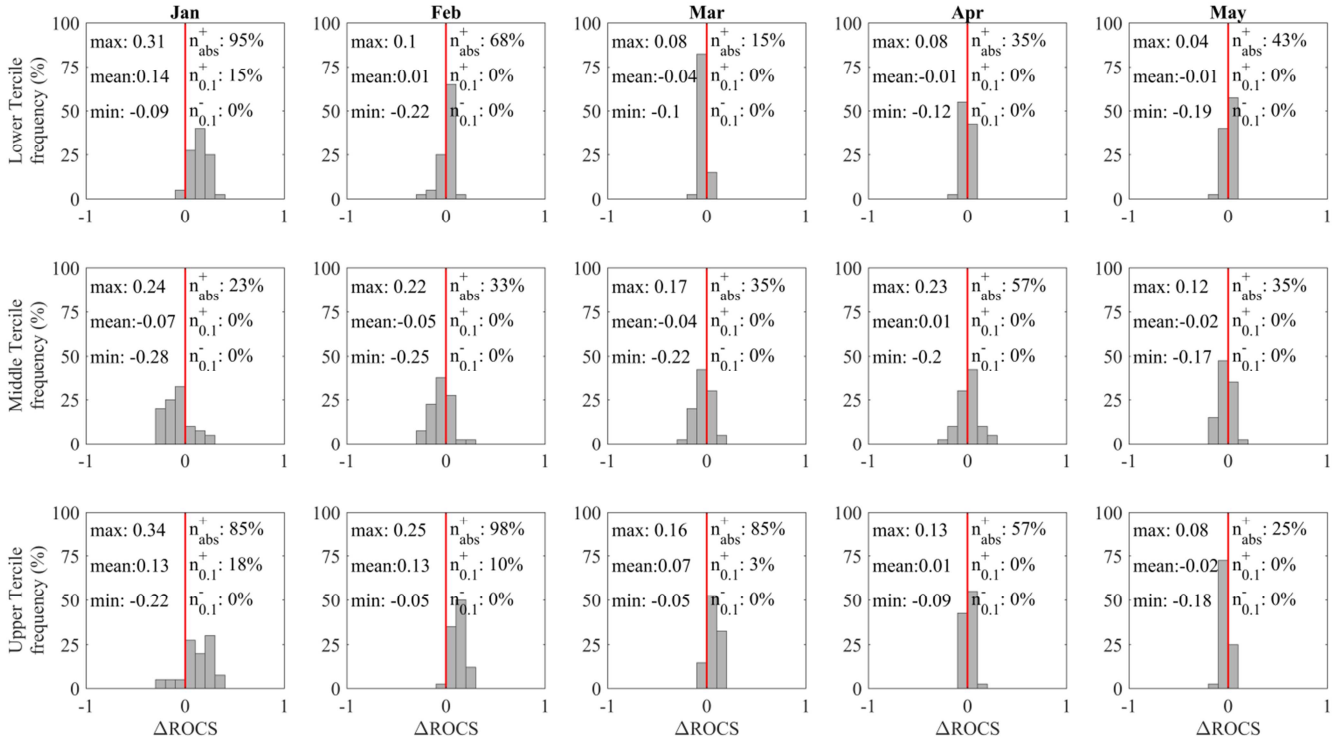

Figure 5. Bootstrapped (N = 10000) ΔROCSS for the lower, middle, and upper terciles between the ME_{hds} and HE for subbasins in the cluster S^3. Each subplot is a histogram of the medians of the bootstrapped validation score's ensembles for each initialisation month. Above the histograms are six related statistics: (left of the red line) the maximum, mean, and minimum of the validation scores shown in the histograms; (right of the red line) percentages of the subbasins where ME_{hds} performed better than HE (n^+_{abs}), the percentage of subbasins where ME_{hds} performed better than HE ($n^+_{0.1}$) at the significance level 0.1, and lastly the percentage of subbasins where ME_{hds} performed worse than HE ($n^-_{0.1}$) at the 0.1 level.

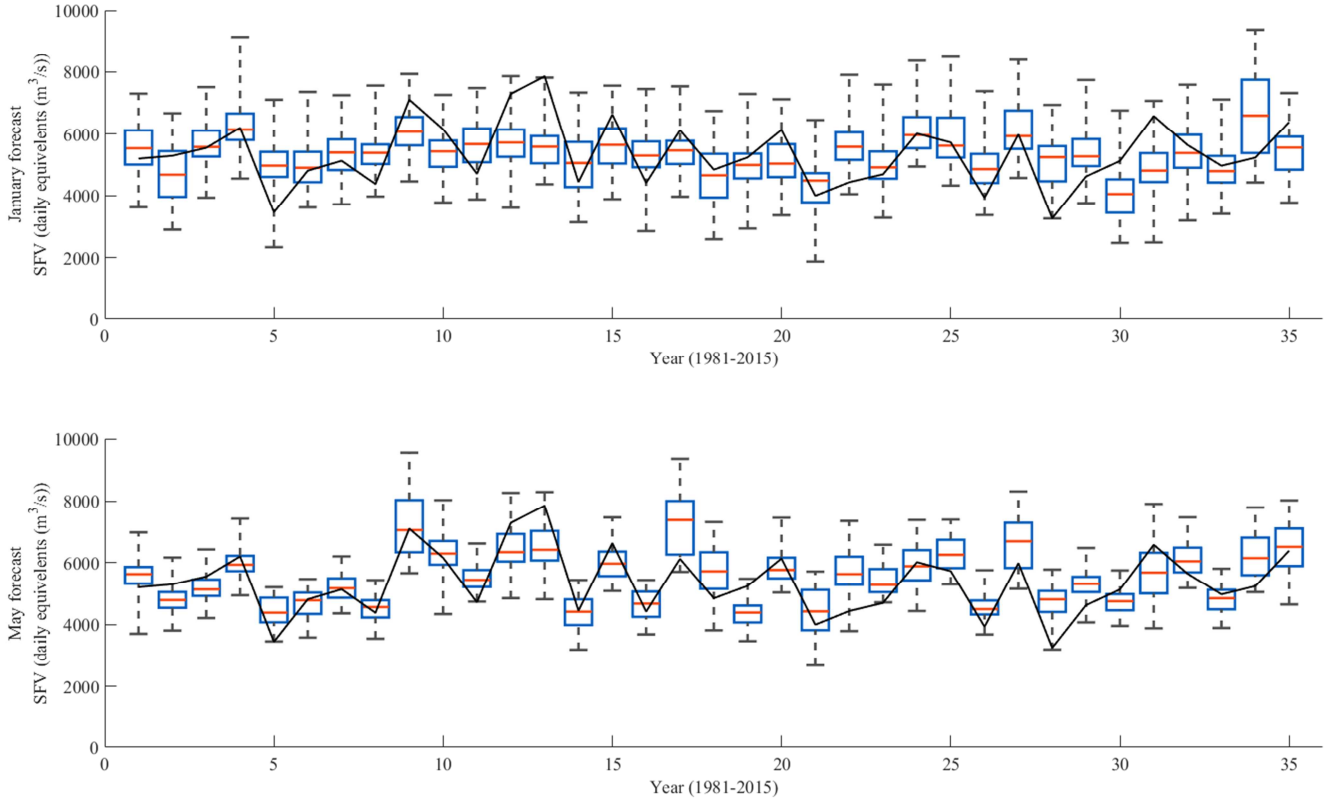

Figure 6. The cross validated hindcasts of the SFV for a subbasin in cluster S^3 made by ME_{hds} (boxplots) together with the observed SFV (black line). The box plots represent the entire forecast ensemble, the red lines represent the ensemble medians, the blue boxes the 25th and 75th percentiles (IQR), and the feelers represent the 0th and 100th percentiles.

Table 4. Bootstrapped (N = 10000) USS and IQRSS for ME_{hds} using HE as a reference. All values that are in bold are statistically significant at the 0.1 level.

		USS					IQRSS				
		Jan	Feb	Mar	Apr	May	Jan	Feb	Mar	Apr	May
S^1	SS	0.21	0.04	0.02	-0.02	0.00	0.01	0.05	-0.02	-0.08	-0.18
	n^+ (%)	80 (16)	64 (0)	56 (8)	48 (4)	52 (8)	56 (24)	72 (20)	40 (8)	16 (4)	16 (4)
S^2	SS	-0.05	0.17	0.07	-0.10	0.17	0.01	0.06	0.05	-0.13	-0.15
	n^+ (%)	48 (4)	80 (4)	60 (4)	40 (0)	68 (4)	52 (36)	68 (40)	64 (16)	16 (4)	12 (0)
S^3	SS	0.05	0.06	0.11	0.07	0.06	0.09	0.08	0.02	-0.05	-0.03
	n^+ (%)	60 (12)	65 (10)	65 (15)	58 (10)	60 (12)	70 (52)	80 (48)	60 (25)	32 (8)	48 (18)

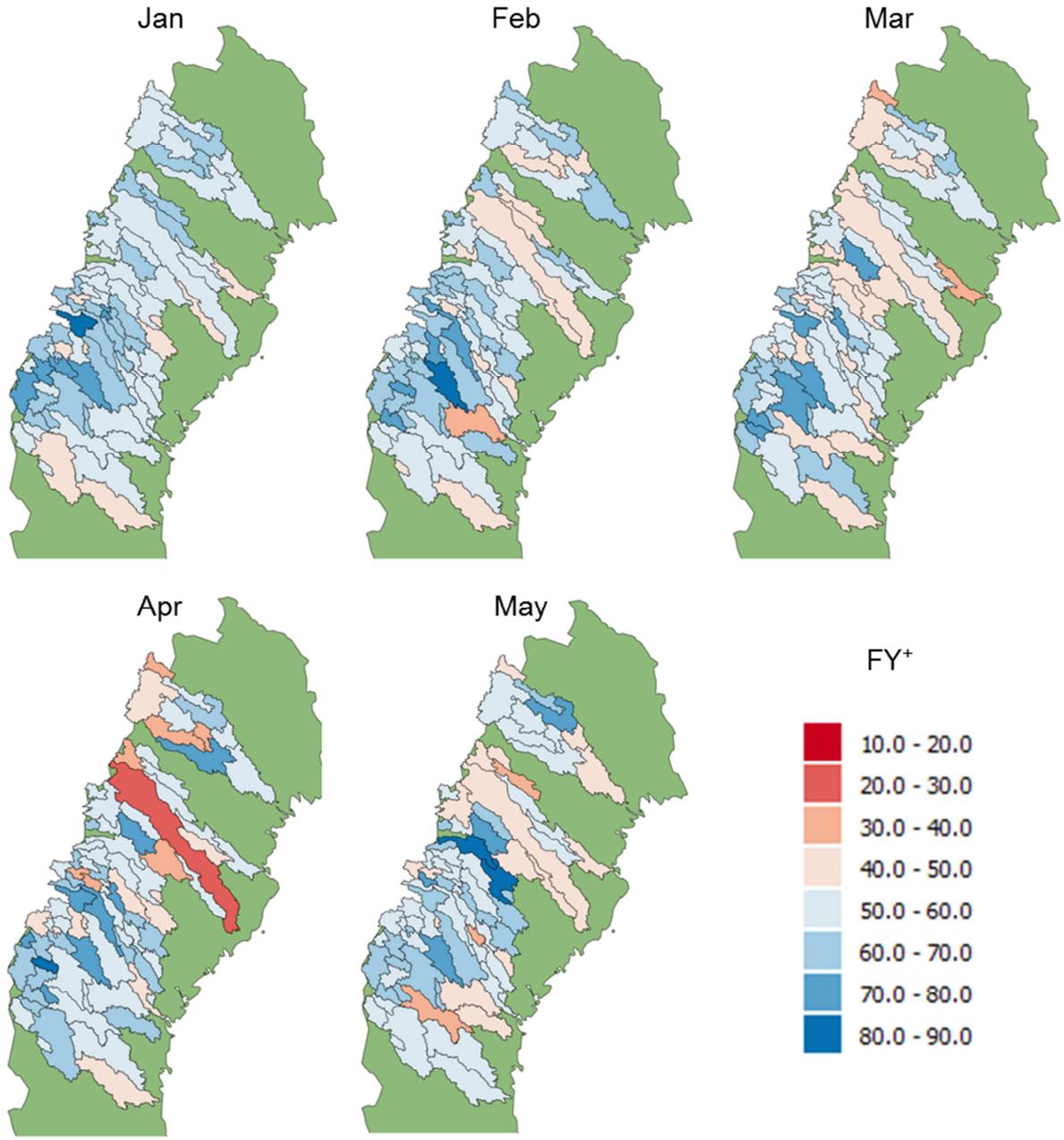

Figure 7. Maps of the median bootstrapped FY^+ values for each of the initialisation dates.