Revision details

The paper has now undergone a major revision taking into consideration the reviewers comments. The revision is summarised below, along with the response to the reviewers.

1. A figure to explain the hindcasts
2. Two figures showing bias and a figure showing reliability was added
3. The discussion of the paper
4. CRPS for was recalculated for the two different ensemble sizes
5. Clearer explanation was added on how the two systems were compared

Response to reviewer 1

Reviewers comment's in blue, our responses in black

In this paper the authors evaluate the added benefit of using a seamless integration (SEAM) of the outputs from ECMWF extended-range ensemble prediction system ENS-ER) and the ECMWF system 4 seasonal forecast system (SYS4) for hydrological applications. The added benefit from this approach is evaluated by comparing the continuous rank probability scores for the outputs from the hydrological model LISFLOOD forced by SYS4, SEAM, and a climatological ensemble (CLIM) over the hindcast period.

The authors find that hydrological hindcasts made using SEAM show better skill, over those made using SYS4, for much of Europe with lead times up to seven weeks. In some areas like the parts of Alps and northern Finland the reverse was true; however these results are uncertain due to the general poor performance of LISFLOOD in these regions. They argue that the increased skill can be attributed to the better initial conditions of the hydrological and meteorological conditions (models are initialised biweekly as opposed to once per month for SYS4) as well as the use of a better atmospheric model in SEAM (the atmospheric model used in SYS4 is locked at the initial version released with system 4 while the one used in ENS-ER is updated regularly). They conclude that the use of SEAM for hydrological forecasting at the seasonal scale has an added value for decision makers given the higher frequency of updates and improved skill, especially at the sub seasonal scale, making the forecasts more actionable.

The topic of this paper is of great interest at the moment considering the increased focus on forecasting at the sub seasonal to seasonal scales in recent years. Although the concept is not new this paper is the first, that I am aware of, that makes an attempt to evaluate a system that utilises current 'off-the-shelf' operational products. The pa per is well written with a good structure and generally clearly formulated, the methods are scientifically sound, and the results are interesting. Additionally, the research presented in this paper is very relevant to the topic of this special issue. In my opinion, the manuscript has a lot of potential for publication in this HESS special issue. However the authors need to clarify some points and revise some statements so that the paper is more easily understood.

General comments:

1) I feel that it is not clear for what periods the study was performed, something which has a bearing on the quality of the results. The authors state that (P5, L132-L135)

"This study focuses on the performance of SYS4 and SEAM over the hindcast period of the operational forecast with a sequence of starting dates over the period 2015-05-14 (the first available date with 11-member hindcast for ENS-ER) to 2016-06-02 producing daily output time series of discharge over the 20-year hindcast period."

The first part of the sentence suggests that the evaluation period is between the dates 2015-05-14 and 2016-06-02 yet the second part says that the hindcast period has a length of 20 years. The next line has a similar mixed message. From the paper I get the general impression that the evaluation is done for the 20 year period so I assume that the issue is to do with how section 2.3 is worded. This should be addressed as there is some confusion in the way that the paragraph (p5, L231-L238) explains it. Further it has implications on the robustness of the results, should the evaluation period be just the 13 months between the aforementioned dates this would give a limited data sample from which to draw the wider reaching conclusions made by the authors. How can the authors know *whether the performance of the different approaches during that period was typical of their general performance?*

The hindcast has 20 years of rerun forecasts, so it is not just one year of integration. The section has now been clarified and we have changed the wording to:

"This study focuses on the performance of SYS4 and SEAM over the hindcast period of the operational forecast. The hindcast of the ensemble forecast is produced twice per week (Mondays and Thursdays) by running an ensemble of 11 members with for that particular day and month, for each of the previous 20 years. The hindcast is run up to 46 days, similar to the ENS-ER. For this experiment, the hindcasts with a sequence of starting dates from 2015-05-14 (the first available date with 11-member hindcast for ENS-ER) to 2016-06-02 were used. This provided 13 monthly starting dates for SYS4 and 111 biweekly starting dates for SEAM with corresponding hindcast set covering all seasons over the previous 20-year period, each with 11 ensemble members. The output was averaged to weekly means before the skill score analysis."

Further, we will add a figure to explain how the hindcasts of the extended-range forecasts are set up.

2) The results show that SEAM has skill over SYS4 in the first 3-8 weeks (Figure2b), mostly concentrated in the first 6 weeks. This would imply that there may be a benefit of merging the two meteorological forecasts before day 46. Did the authors consider this and if not why?

We are not sure if we understand this comment. Fig2b shows that SEAM has more skill than SYS4 for the first 2-3 weeks, but that after that there are some areas where the SYS4 performs better and vice versa. The differences can have many explanations, where geography and altitude plays a part (Fig 3). For those areas where the SYS4 performs better than SEAM, it could as Kean suggests be beneficial to switch to SYS4 earlier than after 46 days. However, that would be interesting from an operational point-of view and is out of scope for this study. A system where you would switch between two systems in an optimal way would have to be carefully calibrated and the effect of switching forecasts would have to be significantly better to justify it. We would rather advocate that SYS4 is used in areas where it is clearly better than SEAM, or as a complement to SEAM. However, we will in future studies dive deeper into the differences in skill between the different forecasts.

Specific comments:

P2, L26: "TSYS4 is also ..." I assume that this is a typo and should read, "SYS4 is also..."

Yes, it was corrected to SYS4.

P7, L224-L225: Although this line is factually correct it appears to contradict the preceding ones. The reader is being told how the low flows during this period caused substantial economic losses due to it affecting inland navigation in the Danube and Rhine basins only then to be told that navigations are regulated during high flows and not low flows. I suggest rewording this or removing this sentence to remove the perceived contradiction or removing this line altogether as it does not add anything significant to the discussion.

The sentence was there to point to the fact that there are no regulated restrictions on the low flow; it is down to the transport companies to make the decision. We agree that it does not add any significant information and the sentence will be deleted in the revised version..

P8, L249-L250: The second part of this line is awkward to read and should be rephrased.

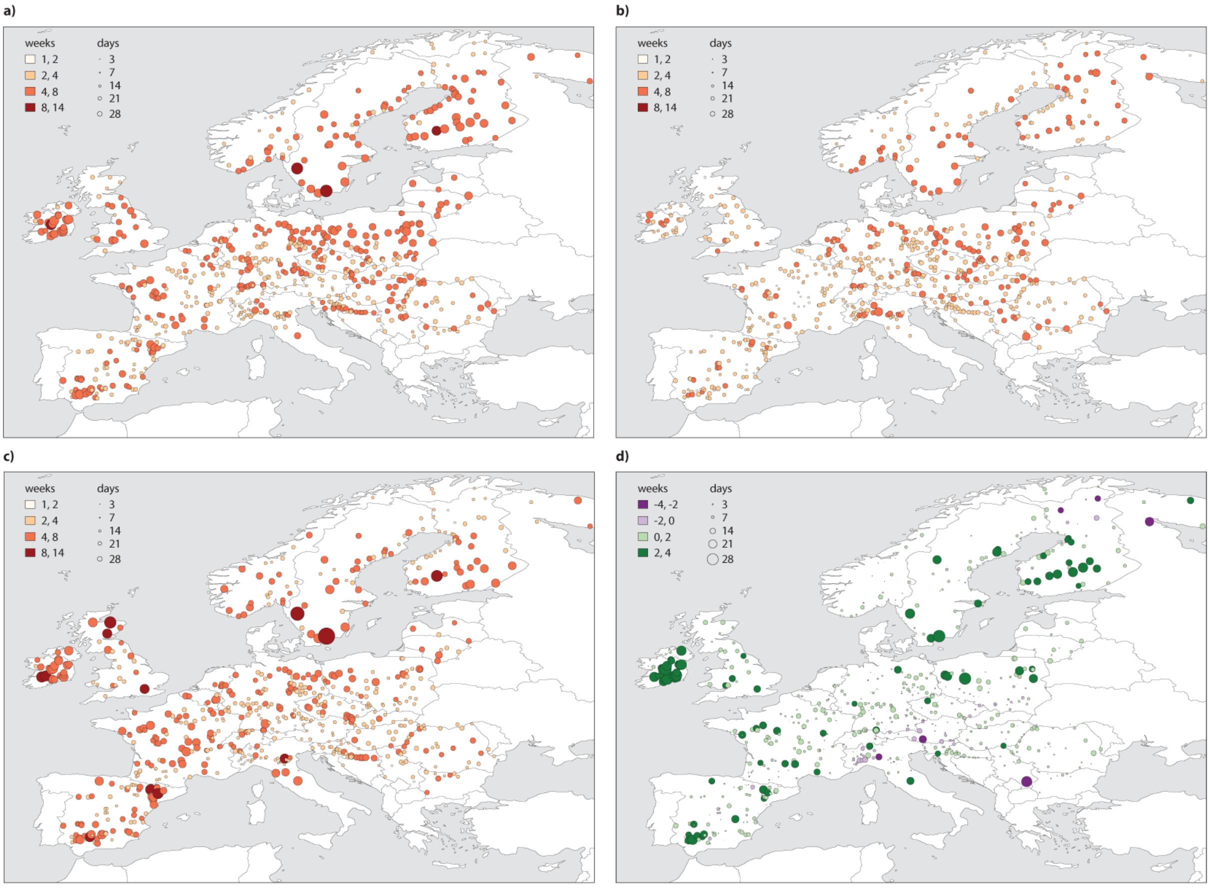
The sentence was rephrased to: “The onset of the second low period was correctly modeled by the SEAM system, whereas the timing of the low flow was missed by SYS4”

P11, L350-L35: I think the reference is - Pappenberger, F., Wetterhall, F., Dutra, E., Di Giuseppe, F., Bogner, K., Alfieri, L., and Cloke, H. L.: Seamless forecasting of extreme events on a global scale, pp. 3–10, Proceedings of H01, IAHS-IAPSO-IASPEI Assembly, Gothenburg, Sweden, July 2013 (IAHS Publ. 359, 2013)

Yes, that is correct; the reference has now been updated.

P16, Caption to figure 3: The last line states, “The dimension of the circles is proportional to the number of days while the color scale refers to progressive weeks.” What do the authors mean by number of days?

The size of the circles are proportional to the number of day of predictability. The circle size was missing in the plot legend that has now been revised. To make the plot more readable we had also colour-coded the circles in broad weekly changes. Clearly, there is a correlation between colour and circle sizes as the darker the colour the larger the symbol dimension. However we found that the colour breaks made the plot more readable. The graphics of the plot has been slightly revised and is as follows:



Response to reviewer 2

B. Klein

The manuscript shows the development and the skill of a seamless hydrological forecasting system from sub-seasonal to seasonal scales. Meteorological forecasts from ENS extended (day 1 – 46) and SYS4 (47 to month 7) are merged by randomly selecting ensemble members of SYS4 after ENS extended ends. The skill analysis shows that most of the skill improvement by using SEAM is due to the more frequent model initializations and the more recent NWP model version of ENS extended. The paper is well written, the methodology and results are nicely presented and compared. The real value of this study is the application of products off the shelf (available operational products). Hence the results can be directly incorporated in real-time operational streamflow forecasting practice. The paper should be foreseen for publication in HESS after minor revisions.

Thank you very much Bastian for the comments. Below are detailed point-to-point answers to your remarks, but we have also taken on board your comments of expanding the discussion and conclusion part on the predictability and limitations of the two systems and will expand the discussion part further.

Comments:

p 2, l 24: Typo, replace TSYS4 with SYS4

The typo has been corrected.

p 2, l 30: please add the forecast length published in the seasonal outlook of EFAS

"...with a lead-time of 7 month." was added to clarify.

p 2, l 119: please add possible drawbacks of selecting a random member of SYS4 (one point was raised p 6 l 192- p7 l 195). Another possible drawback could be that ensemble members are combined originating from complete different climatological conditions day 1 – day 46.

We are aware of this problem and we tried to address it on p6, but will expand on this and discuss the drawbacks further. However, the regimes over Europe can shift quite rapidly and it is not certain that matching the ensembles would increase the skill of the seamless.

p 3, l 89: Are the 5kmx5km grid cells of Lisflood further subdivided in elevation zones?

Yes, they are divided into three sub elevation zones to account for differences in snow accumulation and snowmelt. See more details in the answer to P7 below.

p 4, l 124: Are bias/drift correction methods applied to correct the meteorological forecasts?

No bias correction is applied to the meteorological forecasts

p 5, l 135: the description of the hindcast period used in this study is a little bit confusing due to the mixture of forecast dates (2015-05-14 – 2016-06-02) used to produce the hindcast dataset and the forecast dates of the retrospective forecasts. Please clarify!

One possibility would be probably to add the range of forecast dates. Something like: "...the hindcast data set of SEAM covers the period 1995-05-14 to 2016-06-02..." "...the SYS4 re-forecasts used in this study are initialized each month over the period 1995-05-01 to 2016-06-01..."

Thank you for the suggestion. Also Kean commented on the difficulty of understand the setup of the experiment. We have taken care in explaining the hindcast and experiment setup more in detail. We will also add a figure to explain the setup of the hindcast system.

p 5, l 160: replace SEAS with SYS4

Corrected.

p 6, l 161: Incomplete sentence, I assume: ".... as in SEAM to account for the difference in ensemble size...."

The incomplete sentence was deleted

p 7, l 206: Another option of the poor performance of Lisflood in these regions could be the snow modelling component. In steep orography a 5km x 5km grid is relatively coarse to model snow adequately, are grid cells of Lisflood further subdivided in elevation zones? Please add a comment/discussion of the snow modelling performance of Lisflood.

The snow modelling in LISFLOOD is a degree-day method with elevation zones to further differentiate the snow processes in steep orography. This could explain differences in the model performances if the results were compared with observed runoff. However, the model results are compared with a climatology run using observed precip and temperature, and it is more likely that the poor NWP representation of temperature and precipitation are the culprits.However, the snow modelling component could also play a role in this, and we will add a description and discussion on this to the paper.

"The snow accumulation and snowmelt are further divided into three elevation zones within a grip in LISFLOOD to better account for orographic effects in mountainous regions. However, this increase in sub grid resolution is not likely to be high enough to capture the snow variability during the snow accumulation and snowmelt in mountainous regions. Further, precipitation forecasts have documented biases in steep orography (Haiden et al., 2014).

p 8, l 231: add Figure to the figure number "...Cologne (Figure 4)...".

Corrected

p 8, l 233: I assume 3% of its climatological value is derived from the simulated climatology and not from the observed climatology? Please specify!

Yes, it is correct, we are throughout the paper comparing against modelled climatology. We have taken care to make this very clear wherever this is mentioned in the paper. At the above mentioned passage we have changed the sentence to: "went below the 3% percentile of *the modelled* climatological value". Italics denote the addition

p 8, l 240: It should be mentioned that the second low flow event was hit by the SYS4 forecast initialized 2003-09-01. This signal towards a low flow event is missing in the SEAM forecasts published after 2003-09-01. In SEAM a signal towards an extreme low flow event first appears about 3 days before the begin of the event (forecast date 2015-09-14). I would add the real forecast dates to

Figure 4 and not the forecast dates the hindcast data set is produced. This could be a little bit confusing for a reader not familiar to the hindcast procedure of ENS extended.

Yes, and the example is chosen to illustrate a situation where the SYS4 performed well. We also point to the fact that SYS4 does perform well in this particular case in the discussion. However, the higher frequency of the SEAM would give it an advantage when you are closer to the event, since you would get more detailed information about the timing. The following was added to stress the point: "SYS4 does indicate the second low flow with a longer lead time than SEAM. However, SYS4 misses the timing of the event.

Figure 4 was also improved to show more clearly the forecast dates vs the verification dates.
p 8 Conclusion: I miss a discussion of potential improvements of the presented seamless forecasting system. Are there any ideas how to reduce the higher spread of the CRPSS of SEAM compared to SYS 4 in figure 2 c, d? Probably an improvement of the methodology of the concatenation of the forecasts from the two systems? Please add this aspect to the conclusions.

This is a good point, and still be investigated, however outside the scope of this paper. We will add the following to the Conclusions.

"Future work with the seamless forecasting system is to further explore the limits of predictability to assess the strengths and limitations of the current setup. The assumption that the forecasts can be randomly concatenated would also need to be tested against a system where the forecasts are matched according to their respective climatology."

Another aspect I miss is the conclusion from Figure 2 b):
The improved boundary condition of the first 46 days originating from the more recent model version with a higher resolution doesn't improve the predictability (forecast skill) after day 46.

This is also a good point, and more is to say on the predictability of the two forecast systems. This will however as mentioned above be dealt with in another study, so we are reluctant to speculate too much at this time.

Figure 3: Are all forecast dates used in this analysis? Please add to the caption to be consistent with the caption of Figure 2.

No, in Figure 3 only the first forecast of the month is used to avoid too much the effect of the initial conditions of the hydrological model. This will be clarified in the revised manuscript.

Mike Hardeker

This paper has a good (not new) idea, but is disappointing as it just skims over results without proper analysis. It currently does not have a proper scientific discussion and reads like it was rushed. In addition, the paper seems to have written for a different journal, it is extremely short (which is good in theory), but simply lacks depth and proper analysis. Results are not properly explained and leave many questions. This is best illustrated by the use of a single score, which does only measure one property of an ensemble forecast - I would have at least expected some de-compositions.

Thanks for your comments. This paper is purposely short as we intended to showcase only one result: how much is the gain in using a seamless forecast system instead of the seasonal forecast. As you also point out this idea is not new and has been referred to in many other papers. However, to our knowledge, this is the first paper that quantifies what is the effective gain in an operational system in weeks of predictability. As such, this paper tries to diagnose the advantage of a concatenated system against the exclusive use of System-4 which is often the preferred choices.

Is it true that this paper leave questions open. The most urgent one in our view is what happens if someone has only access to the seasonal system? This is quite common as seasonal forecast is freely available as opposite to the ENS forecast. There are ways to improve the predictability of the seasonal forecast for example by applying the finding of this paper. We are preparing another work looking specifically to this aspect and exploring in details the sub-seasonal to seasonal predictability.

We want to keep this work as much as possible focused on this single question, however, we agree that other diagnostic could be added and we will extend the results including bias and reliability in the revised version, and also extend the discussion.

Detailed comments: Acronym ENS-ER appears in introduction first and needs to be defined in introduction not only abstract. I could not find that acronym on ECWMF's websites which makes me wonder what the authors have actually used.

We have added the explanation of the acronym to the text. The extended range ensemble prediction system (ENS-ER) refers to the bi-weekly 46 days extension of the otherwise daily ensemble prediction. ENS is the official acronym for the ensemble forecast, and we added ER to distinguish this from the normal ENS. Even if this is not an "official" ECMWF naming convention, we found this to be a useful acronym for this paper

The introduction defeats most of the paper. I clearly states: "This implies that the skill of SYS4 is lower relative to ENS-ER in the overlapping first six weeks (Di Giuseppe et al., 2013)", which is obviously a result that has been already published by one of the authors earlier.

The idea of the paper is not to assess the fact that the ENS-ER is a better forecast, it is how much better it is and whether it can be used in together with SYS4 to create a seamless forecast, which is updated more frequently than the seasonal forecast. This is important information for any user of hydrological seasonal forecasts, as was also pointed out by the other reviewers.

Also in (Di Giuseppe et al., 2013) we assumed that the ENS-ER in the overlapping first six weeks was better that than System-4 and then went on doing other analysis. The fact that this assumption has always been accepted without questioning reinforces the idea of this paper, which tries to quantify those statements. Di Giuseppe et al., 2013 looks at forecast calibration for the purposes of generating a malaria early warning system. There is almost no overlap with what done here apart from the use of the ENS long range forecast.

L34 it is unclear why the extension leads to benefit. Point (ii) - that has been possible before, what is better and why? There are no references stated for the hypothesis listed in (i) to (iii) - a more detailed in depth discussion and reasoning (or supporting results) are needed.

Regarding point 2:, the previous hindcast of the monthly extension was only 5 member and up to day 32. The new system with 11 members and lead-time 46 days is much more useful than the previous system, therefore the possibilities of carrying out pre-and post-processing has greatly increased.

The first point is related to the fact that we can now extend the ensemble forecast more in time than was possible previously, and that we can issue seasonal forecasts with higher frequency than before, given that the method of concatenation ito to a seamless forecast is working well.

The third point is a bit more speculative, but a decision support system for more products that was previously available would be possible and feasible to implement. The examples of benefits are given in the below statement.

We will expand on these three points and support them with references.

"The extended lead time provided by running EFAS forced by weather prediction across different time scales could potentially provide added benefit in terms of very early planning, for example for agriculture, energy and transport sectors as well as water resources management." - where is the evidence for that statement? references? Studies - this unsubstantiated and symptomatic for the rest of the paper - many claims or statements which are not backed up.

The statement is quite modest. We are simply saying the availability of a skilful forecast X days ahead is more useful than a skilful forecast provided Y days ahead if X>Y. We would imagine this to be uncontroversial. If in some sectoral application there is only need for Y days forecast, then the X days information can be easily disregarded. Forecasts are used in many applications, and we will substantiate that with more references to such studies.

"often model implementation is segmented for practical reasons. Still major efforts have been made to create unified systems" - it is completely unclear what is meant - clarify

As the reviewer points out our introduction is quite long as we were very comprehensive in highlighting the context from which this paper was generated. We have explained that the various weather prediction systems have been developed from requirements that have been added in time as weather forecast has improved in terms of predictability. This has led to fragmented systems. This fragmentation is somehow not intentional, however practical. Some institutions have gone all the way to rewrite their model (UKMET office) so that this could be used at all time scales. These systems could provide possibly a better tool for predictability studies. However, this work does not try to quantify predictability per-se but to put a predictability length to one of the most used system in the world, given that one takes what is available from the shelf. If the reviewer is searching for a theoretical study this is not the right paper. As the two other reviewers have pointed out this paper has value as it analyses in the specific a very well used system even if the results validity are then limited to that particular system.

"Similarly, the UK Met Office has in the past twenty-five years worked to create a unified model that could work across all scales (Brown et al., 2012). Also the climate community has moved in the same direction. For example, the EC-Earth project shows that a bridge can be made between weather, seasonal forecasting and beyond (Hazeleger et al., 2010, 2012)." this is not relevant for the paper. I am unsure what point the au-thors are trying to make with respect to the hypothesis tested in this paper.

This sentence is instead quite relevant as it compares our concatenation approach to another approach (creating a unified model) that exists even if it is not used in this paper. We believe it is part of the bibliographic review process in the introduction to acknowledge what is available even if is not used.

Introduction needs significant shortening.

Sorry but we disagree as we find our introduction quite a nice historical overview of the conception, the designs and the different approaches followed for the practical implementations of seamless forecasts.

"avoiding the complications of new developments while generating forecast products to meet different types of users (Pappenberger et al., 2013)." Pappenberger is clearly wrong - one will always need different products for different applications.

We are not arguing that a specific users do not need to tailor the product to meet their needs, just that you can achieve quite a lot with already existing information. The two systems, the seasonal and the extended range are both worth using to a larger extent than they currently are, and they are readily available. The tailoring towards your own needs is necessary for any application as you clearly state, and that is exactly what we are doing when we are using the meteorological forecast to force a hydrological forecast.

"diverge over time, only re-converging when the seasonal system" That assumes that the seasonal system is very close to the system from which it is derived from. I just googled ECMWF System 5 and it seems to come from an older model cycle, hence this statement is clearly incorrect

We agree, the statement is too strong, the systems never completely converge, the gap in model cycles are shortened with a new release of a seasonal forecast. We have changed the wording to:

"One important consequence of this difference in design is that, for example, the much more frequent updates to the extended range compared to the seasonal system at ECMWF, imply that the bias characteristics of the two systems diverge over time, only closing when the seasonal system is updated."

"final products should be provided in terms of anomalies calculated against the model climate" that assumes that the model universe behaves similarly to the real universe in terms of anomalies - can the authors provide any prove and evidence?

Yes, the EFAS system behaves well in terms of issuing forecasts in comparison with the model climatology. It is not perfect, so is no system. EFAS has been calibrated against observations where they are available, and the performance is generally good. The praxis of EFAS is to compare against its water balance, this is the standard procedure. We can be clearer in the references to previous studies regarding this.

This argument here is rather that the concatenation itself needs to be taken care of since it is likely to create a bias when the two systems are combined. We have added a sentence to point tos this argument.

"What is the gain of using a more recent model version in the first 46 days provided by the use of the ENS-ER?" I don't understand that question cause according to the authors this has been already answered in a paper cited by the authors themselves, (Di Giuseppe et al., 2013). It demonstrate that the paper currently only presents a very very incremental step.

In Di Giuseppe et al 2013 we assumed that a seamless system would have been better than the seasonal forecast, however we never proved it neither we looked at the differences with system-4.In this paper we are actually proving what is the benefit of using a seamless system.

It is unclear how the authors come to 786 reference points - how have they been choosen - the claims made by the authors are not substantiated by the results presented. Can the authors please add the analysis which lead to those points? this is a clear example where the paper has been cutting corners rather than explaining properly what has been done.

This will be more clearly explained. We also apologise for an error, the final number of reference points were 679, not 786 as originally stated. The reference are the EFAS outlets from the several sub catchments in the domain. They were chosen as representative points for the performance, and were the points that were used for the operational calibration of EFAS. We will state this more clearly in the paper. We will also add references to the literature where more detail can be found.

However, we do not understand the comment on why the claims are not substantiated by the results? In fact, we could have choses a random number of points, or all of them, and the results would still have been valid as long as we are comparing against a modelled water balance. The selection of these particular ones was to have a reasonable number of points with a good geographical spread to assess the performance of the system.

" (referred to as tuning in the NWP nomenclature)" This is a hydrology journal, why do you explain that?

The journal is read by both meteorologists and hydrologists. Often the two communities use different nomenclature for the same process. We do not think there is any harm to explicitly clarify this aspect for the benefit of a vaster reader audience.

"Using the WB run as proxy observation simplifies the interpretation of the skill scores as it avoids the complication of having to assess the bias against observed discharge." This maybe convenient to do, but then the analysis could have been done against all grid points or far more (700 is pretty low given the size of that Grid). The authors need to elaborate on the limitations this analysis places on the results of the study. I am also thoroughly confused, the authors said that they had real observations for the calibration. I would expect at least some analysis against those real observations. Far more detail needs to be provided.

To answer the first comment on the number of points used for the assessment of the system. The total number of points at which discharge is calculated over all of EFAS is 38452. We could have calculated the performance on each of these points, and we routinely do that as part of our performance. However, since they would in many cases be highly correlated (points along the same river will behave similar), a sub-sampling was made to represent the performance over the entire domain. This was a conscious decision to simplify the calculations and to avoid too correlated skill scores, as independent sampling as possible. We consider the selection good enough to represent the performance of the system and do not see the reason to increase the number of points.

The second question regarding why we did not include the observational data has been discussed in the paper. The EFAS system is covering the entire European continent and can as such not be perfectly calibrated everywhere, especially not on a 5km grid. The observations are alos not available for the full hindcast period at each location.

The water balance run, which is the model performance using observed precipitation and temperature, are a proxy for observations, and is what we chose to compare the performance of the models against. The benefits of using the water balance is to avoid observational errors and also to mimic the performance of the operational EFAS forecast system, where the forecasts are also compared with the water balance rather than observations. Since we are comparing the two forecasting systems and not trying to assess the total skill of EFAS, the use of the water balance run is justified. We understand that this was not fully explained in the paper, and will add this to the discussion.

"The hindcast period can together with observations be employed to calibrate the forecast in an operational setting (Di Giuseppe et al., 2013)." I am unsure about what the authors mean with that statement and find the reference strange and forced (deliberate self citation?). Can the authors please cite references from others too?

This paper is cited as an example of a correction that can be calculated from the hindcast set and then applied to the forecast. The methodology developed in Di Giuseppe et al 2013 is quite complex as it was designed to correct a precipitation systematic southerly shift in the west African monsoon. However, the calibration was implemented for the exact same system used here, i.e. a seamless concatenation of the ENS-ER and system-4. For this reason, we thought it was a well suited reference. However, following the suggestion we have added other two well-known work for bias correction.

Figure 1 is unclear - how do different ensemble number play a role. Did you only merge 11?

We have added a new schematic, which explains in better details how the hindcast set from the seamless, is constructed. Since there are only 11 hindcasts of ENS-ER, only 11 could be merged with the hindcast of the seasonal forecast.

2.3. Experimental set-up - you are comparing apples with pears. One system has clearly a much larger sample size and the authors do not explain how the adjust for that fact. Results cannot be robust unless this is taken into account. Please revise your method thoroughly.

This is taken into account in the analysis (see figure 2) where we compare only the hindcast from the first of the month from SEAM with the seasonal forecasts, therefore not using all forecasts from SEAM. Same as in figure 3, where we only use the first forecast of the month from SEAM in comparison with SEAS. We thought it useful to show the performance over the entire period in figure 2a, therefore it was added. We will make the this clearer in the description of the methodology..

CRPS is equalised by randomly drawing from the distributions - that is at odds with the statistical literature. Check for example this presentation: http://empslocal.ex.ac.uk/people/staff/ferro/Presentations/ems2013ferro-fair.pdf

The random drawing of members from the SYS 4 distribution would not induce a large error, since the members are interchangeable. However, this will be corrected in the revised version where we will instead use the 15-members ensemble from the SYS4 and then corrected for the variation in ensemble size, as suggested by Ferro et al, 2008.

The authors need to present more scores and analysis. They talk explicitly about droughts in the introduction - this scores does not analyse. To understand skill, one needs to look at least at the decompositions of the CRPS. The analysis needs to be extended significantly and far better discussed. "then some points show a benefit of using the SYS4 instead of SEAM." - why? explain

We mention droughts and low flows as possibly uses for a seasonal forecasting system; we do not state that we will look into to it in this paper. We will add reliability and bias to the analysis, as stated earlier. The better performance of SYS4 at certain locations is not strange, we do not expect SYS4 to be outperformed at all locations. However, the exact reasons for the better performance for each location is beyond the scope of this paper and will be more looked at in later studies.

"In the above example, a decision maker would have to make a decision based on a forecast that was issued 2.5 weeks earlier, which would inherently make the decision more uncertain if you only had the seasonal forecast. With the seamless system available a decision maker would gain the same early indication of a hazardous event and also have the benefit of frequent updates." Can the authors please test their hypothesis and provide prove for such unsubstantiated statements? where is the social scientific evidence?

In this statement we are just are just stating that a decision substantiated by the availability of more detailed information is more robust and less of guesswork. We refrain from any speculations as to what the implications are for the human intervention in the forecast process, merely that the forecast is substantially better and more importantly, more frequently updated.
The only situation we can foresee in which this statement could be misleading is if the seamless forecast were not as accurate as the seasonal forecast, in which case a bad information might be worse than no information at all. As this is not the case and, it has been clearly proven throughout the paper, we do not see how this statement can be considered "unsubstantiated" as it is just driven by common sense.

I do not understand the point of section 3.3. - it presents a single case and then makes some wild statements. Please assemble a larger number of cases or simply cut

Section 3.3 does not claim to provide any statistical significance of the quality of SEAM against SYS4. This is done in the sections before. Here we have made a practical example for a case studies looking at what the more timely information provided by SEAM could imply in a decision making context. We believe the discussion that follow from figure 4 is not “wild” instead tries to explains in, admittedly, a simplified scenarios, which kind of product improvements could be achieved given the availability of the seamless system.

the analysis overall falls short for more details. It simply skims over results without really going into them and properly analysing them. Many hydro aspects are ignore. Please explain how your results are driven by spatial variations of the weather forecasts.

We understand that the suggestion is to perform a full sensitivity study of the presented results looking at the predictability arising from weather regimes /patterns. Looking at the hydrological predictability at different time scales as driven by weather is certainly an extremely worth matter, however it would require an analysis that is outside the scope of this paper and we see this more like the subject of upcoming studies.

Conclusions are not comprehensive enough and a proper scientific discussion is missing

Section 3 is named “result and discussion”. As a matter of style preference we have decided to detailed discuss our results in this part of the paper. In the conclusions we only highlight the main novel aspects of the paper without repeating the discussion which takes place in session 3.
As the paper aims at answering one

References

Di Giuseppe, F., Molteni, F., and Tompkins, A. M. (2013) A rainfall calibration methodology for impacts modelling based on spatial mapping, Quarterly Journal of the Royal Meteorological Society, 139, 1389–1401.

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

Manuscript prepared for Hydrol. Earth Syst. Sci.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 13 February 2018

The benefit of seamless forecasts for hydrological predictions over Europe

Fredrik Wetterhall[1] and Francesca Di Giuseppe[1]

[1]European Centre for Medium-range Weather Forecasts, Shinfield Park, Reading ,UK

Correspondence to: Fredrik wetterhall: fredrik.wetterhall@ecmwf.int

Abstract. Two different systems provide long range forecasts at ECMWF. On the sub-seasonal ~~range~~ time scale, ECMWF issues an extended-range ensemble prediction system (ENS-ER) which runs a 46-day forecast integration issued twice weekly. On longer time scales the current seasonal forecasting system (SYS4) produces a 7-month outlook starting from the first of each month. SYS4 uses an older model version and has lower spatial and temporal resolution than ENS-ER, which is issued with the current operational ensemble forecasting system. Given the substantial differences between the ENS-ER and the SYS4 configurations and the difficulties of creating a seamless integration, applications that rely on weather forcing as input such as the European Flood Awareness System (EFAS) often follow the route of the creation of two separate systems for different forecast horizons. This study evaluates the benefit of a ~~sub-seasonal to seasonal hydrological application using a seamless forecasting system obtained from the concatenation~~ seamless integration of the two ~~long-range forecasting systems at ECMWF. Two attributes of were found crucial for the better performance of the~~ systems for hydrological applications and shows that the benefit of the new seamless system when compared to the seasonal forecast ~~;~~ can be attributed to (1) the use of a more recent model version in the sub-seasonal range (first 46 days~~and the gain in skill due to the frequent forecast updates~~) and (2) the much more frequent updates of the meteorological forecast.

1 Introduction

ECMWF produces a range of forecasts, among them a 10 day deterministic high resolution forecast (HRES) ~~,~~ and a lower resolution 15-day 51 member ensemble prediction system (ENS) ~~forecast~~ that is extended to 46 days twice weekly (Mondays and Thursdays at 00UTC~~) (Vitart et al., 2008), and an ensemble seasonal forecast system System-4~~ ; Vitart et al. 2008). In this paper we refer to the extended ENS as ENS-ER. On longer time ranges ECMWF issues a seasonal ensemble forecast system (SYS4), operational since November 2011. SYS4 issues a 7-month prediction (extended to 13 months four times a year) ~~prediction~~ once every month ~~(Molteni and co authors). The ENS extended range (~~(Molteni et al., 2011). The ENS-ER ~~)~~ forecast system benefits from frequent updates [1] ~~that are made to~~ of the model physics and data assimilation system (Vitart et al., 2008). ECMWF releases official model updates on average 2-3 times a year which typically include new improved schemes for physical processes, better use of observations and their assimilation and sometimes increase in model resolution. The seasonal forecast has a ~~lower-resolution and~~ lower resolution, is an older model version than ~~ENS and is~~ ENS-ER and is also updated much less frequently. ~~The~~ This implies that the skill of ~~SYS4~~ the seasonal forecasting system is lower relative to ENS-ER in the overlapping first six weeks~~(Di Giuseppe et al., 2013)~~.

Applications that use numerical weather predictions as forcing, such as the operational European Flood Awareness System (EFAS (Thielen et al., 2009; Bartholmes et al., 2009; Smith et al., 2016)), are often designed for a specific purpose. EFAS has since the start focused on early warning of floods in the medium-range forecast horizon, ~~3-15~~ typically up to 15 days. Recently, a seasonal hydrological outlook forced by SYS4 was implemented operationally ~~.~~ with a lead-time of 7 months (Arnal et al., 2017).

This extension to the monthly and seasonal scales is potentially very useful in order to; (i) produce products which extend the previous forecast horizon; (ii) benefit from hindcasts for pre- and post-processing to produce output of higher quality (e.g. model based return periods); and (iii) design completely new early warning frameworks complementing the existing ones. The extended lead time provided by running EFAS forced by ~~sub-seasonal to seasonal forecasts~~ weather prediction across different time scales could potentially provide added benefit in terms of very early planning, for example for agriculture, energy Bazile et al. (2017) and transport sectors Meißner et al. (2017) as well as water resources management Sene et al. (2018). Such a forecast system would be a first step to close the identified gap between hydrological forecasts on the medium (up to 15 days) and seasonal range (White et al., 2017). These extended range systems may not be able to capture extremes of short-lived events like floods, but they are able to detect anomalous conditions on longer lead times, such as low flows (Meißner et al., 2017) and droughts (Dutra et al., 2014).

The concept of seamless ~~forecast~~ forecasts was first introduced by Palmer and Webster (1993). Palmer et al. (2008) formally expanded the idea showing how short-lived phenomena under certain conditions may persist and increase predictability at longer time scales. Since then the concept of a unified or seamless framework for weather and climate prediction has been vastly debated (Hurrell et al., 2009; Brunet et al., 2010). However as noticed by Hoskins (2013) in his seminal paper, while "the atmosphere knows no barriers in time-scales", often model implementation is segmented for

[1] ~~ECMWF releases official model updates on average 2-3 times a year. These can include new improved schemes for physical processes, better use of observations and their assimilation, increase in model resolutions.~~

practical reasons. Still, major efforts have been made to create unified systems. Indeed, the ENS-ER was the first attempt to create a seamless extension ~~to~~ of the ECMWF medium-range forecast to the monthly scales(Vitart et al., 2008). Similarly~~the UK Metoffice~~ , the UK Met Office has in the past twenty-five years worked to create a unified model that could work across all time scales (Brown et al., 2012). Also the climate community has moved in the same direction. For example, the EC-Earth project shows that a bridge can be made between weather, seasonal forecasting and beyond ~~and can stretch as far as Earth-system modeling~~ (Hazeleger et al., 2010, 2012).

The latter projects went all the way to create new systems starting from existing components and were therefore costly and time demanding. In contrast, a practical and simpler approach could be taken. The seamless idea could be translated into the simple concatenation of "the best" forecast at each ~~lead-times~~lead-time. The clear advantage of this off-the shelf seamless prediction conversion is that it utilizes products that are already in place, thereby avoiding the complications of new developments while generating forecast products to meet different types of users (Pappenberger et al., 2013). There is however an underlying complexity in this simplification; the substantial difference in design between the various forecasting systems makes the concatenation ~~a task~~ technically difficult. As systems are designed for different users they often have non-matching temporal and spatial resolutions, different hindcast span and different ensemble sizes. One important consequence of this difference in design is that, for example, the much more frequent updates to the extended range compared to the seasonal system at ECMWF, ~~imply~~ implies that the bias characteristics of the two systems diverge over time, only ~~re-converging~~ closing this gap when the seasonal system is updated (Di Giuseppe et al., 2013). Then model outputs either need to be bias-corrected to be useful forcing to drive sectoral models such as EFAS, or that final products should be provided in terms of anomalies ~~compared to~~ calculated against the model climate, taken into consideration the bias of the seamless forecast system. In both cases the seamless system needs to account for the use of the hindcast dataset and the application of some bias correction algorithm. In return, the advantage is in the gain in ~~skills~~ skill and the extension of the lead-time.

In this work the ~~added~~ benefit of a seamless ~~hydrometeorological~~ hydro-meteorological system was tested for ~~sub-seasonal to seasonal~~ a span of time ranges from 1 week to 6 months for stream flow forecasts over the European domain using the EFAS system. The aim was to test whether integrating ~~medium-monthly range forecasts with and~~ medium-range forecasts with seasonal prediction contributes to enhance hydrological predictability on the seasonal scale. Specifically, the questions addressed were: What is the gain ~~of~~ in terms of hydrological forecasting of using a more recent model version in the first ~~6 weeks~~ 46 days provided by the use of the ENS-ER? ~~what~~ What is the skill gain provided by having more frequent forecast updates?

2 ~~Method~~Methods

2.1 Hydrological model system

The hydrometeorological system used in this study was the European Flood Awareness System (EFAS ~~) (Thielen et al., 2009; Bartholmes et al., 2009; Smith et al., 2016)~~Thielen et al. 2009; Bartholmes et al. 2009; Smith et al. 20 EFAS is an operational early warning system covering most of the European domain and has been run operationally since October 2012 as part of the COPERNICUS Emergency Management Service (CEMS). The hydrological component of EFAS is the distributed rainfall-runoff model LISFLOOD (De Roo et al., 2000; Van Der Knijff et al., 2010; Burek et al., 2013). LISFLOOD calculates the main hydrological processes on sub-daily and daily time-scales that generate runoff, such as soil and ground water interactions, for each grid cell. In the operational setup EFAS covers most of Europe on a 5x5 km equal-area grid. The runoff is transformed through a routing scheme to estimate the river discharge at each grid cell along the river network. The routing scheme also takes into account water retention in lakes and reservoirs. This study will concentrate on the forecast of river discharge ~~, and more specifically on 786 reference points on the river network across the EFAS domain. These points were chosen as they are the ones that have good historical observations and has been used to calibrate the model and represent both larger and smaller rivers~~at the outlets of the sub basins of the river network that were used for calibration of the current EFAS system (Smith et al., 2016; Zajac and Bianchi., 2013). The total number of outlets used were 679, and they represent river basins of all sizes and characteristics across the EFAS domain.

In its operational implementation ~~LISFLOOD has undergone an~~ the latest calibration (referred to as tuning in the NWP nomenclature) ~~using~~ of LISFLOOD used an observational dataset of meteorological forcing data (precipitation and temperature) and observed discharge ~~which covers~~ covering the model domain over the period ~~1990-onwards~~1990-2013 (Smith et al., 2016; Zajac and Bianchi., 2013). The meteorological dataset comprises more than 5000 synoptic stations that have been interpolated to a 5x5 km Lambert azimuthal equal-area projection (Ntegeka et al., 2013). The high resolution gridded observation of precipitation and temperature were used for the calibration of LISFLOOD. The observational meteorological dataset was also used to generate a reference modeled climatology of discharge (hereafter called water balance, WB) which is used as; (i) initial conditions for the operational forecast and hindcasts and (ii) reference model run to assess the performance of the forecasts. Using the WB run as proxy observation simplifies the interpretation of the skill scores as it avoids the complication of having to assess the bias against observed discharge. The purpose of this paper is rather to assess the skill of the two forecasts used for forecasts rather than the total skill of the forecasting system.

2.2 Seamless integration of meteorological forcing data

Twice weekly every Every Monday and Thursday , the ENS-Extended Range (ENS-ER) forecast at ECMWF issues a 46-days forecast integration ECMWF issues an extended-range ensemble forecast (Figure 2ENS-ER) by continuing the integration time beyond day 15 up to day 46, with a lower-resolution model (Figure 1, Table 1). Each ENS-ER integration comes with an 11-member hindcast set produced for the same dates over the previous 20 years. This hindcast set provides identical integrations as the current operational forecast with the difference that ERA-Interim reanalysis (ERAI; Dee et al. (2011)) and ERAI land reanalysis (Balsamo et al., 2015) is are used to provide the initial conditionsfor the hindcast, whereas the operational ensemble forecast uses the operational analysis. The hindcast period can data together with observations be employed can be used in many applciations, for example to calibrate the forecast in an operational setting (Di Giuseppe et al., 2013). Thus, twice every week a set of 21 years of 46-days ensemble predictions is available using the same forecast system.

The operational seasonal forecast (SYS4) issues a new forecast at the beginning of each month with a lead-time up to 7 months, four times a year extended to 13 months (Figure 2). SYS4 has a hindcast consisting of 30 years started at each month and consisting of 15 members. The new seamless forecasting system (hereafter called SEAM) was created by concatenating each ENS-ER ensemble member with a randomly selected SYS4 ensemble member at day 46, which is the last day of the ENS-ER (Figure 2). SEAM benefits from the frequents updates of the ENS-ER and has the seven months horizon of the seasonal system. As

Since the two systems have different resolutions (table 1) the horizontal resolution was homogenized to the 5x5 km equal-area grid through a mass-conserving interpolation for precipitation and a bilinear for temperature before it was used as input to the hydrological model in EFAS. The time step was reduced to daily by averaging (accumulating for precipitation and evapotranspiration) the three hourly outputs of the ENS-ER and the six hourly outputs of SYS-4. Since the ENS-ER has a reduced hindcast (20 years) and number of members (11), SEAM has the same number of members and hindcast period. Note that in real-time mode, a full 51-member SEAM is possible. The technical details of the forecast and the hindcast used in this experiment are presented in table 1. For simplicity SYS4 and SEAM will from now on refer to the full hydro-meteorological integrations for the remainder of this paper.

2.3 Experimental set-up

This study focused focuses on the performance of SYS4 and SEAM over the hindcast period hindcasts of the operational forecastwith a sequence of starting dates over the period . The hindcasts starting from 2015-05-14 [1] to (the first available date with 11-member hindcast for ENS-ER) to 2016-06-02 producing daily output time series of discharge over the 20-year hindcast period. The output was averaged to weekly means before the skill score analysiswere used used as input to the full EFAS

[1] The first available date with 11-member hindcast for ENS-ER)

modeling chain. This provided 13 monthly starting dates for SYS4 and 111 biweekly starting dates for SEAM with a corresponding hindcast set covering all seasons over the previous 20-year period, each with 15 and 11 ensemble members ~~,~~ respectively (Fig. 1). The output was averaged to weekly means before the skill score analysis. Since the starting dates of the SEAM and SYS4 were not always in sync (the starting date of the SYS4 integrations are only sometimes on a Monday or Wednesday), it is impossible to do a completely like-for like comparison since the validation periods would be slightly different. However, this error will be random and given the sample size (260 and 2220) it was not considered to have a big impact on the results.

SEAM was ~~verified~~ validated against the runs with SYS4 to assess the added value of the merged forecast. Further, both model systems were compared against a climatological benchmark simulation (hereafter called CLIM). CLIM was constructed by forcing the LISFLOOD with ~~11~~ 15 randomly selected time series of observed meteorological forcing from the period 1990-2014, excluding the modeled year. CLIM has the advantage of having the same initial conditions as the SYS4 and ~~SEAS~~ SEAM hindcasts, but has no expected predictive skill beyond the horizon of the initial conditions. The advantage of CLIM is that in theory it has near perfect reliability ~~as~~ with regards to the WB runs since it is produced with the same unbiased forcing data. It should therefore score better or equal as the hindcasts as predictor on time ranges beyond their respective limits of predictability.

2.4 Score metrics

The performance of the two ~~forecasts were quantified using the~~ forecast systems was compared against modeled discharge using observations at the 679 sub basin outlets using deterministic and probabilistic scores. We refer to this run as observations in the paper. The scores used were the continuous ranked probability score (CRPS~~, (Hersbach, 2000)) applied to the modeled discharge over the 786 reference points~~; Hersbach 2000), mean relative error (MRE) and forecast reliability. CRPS is a common tool to ~~verify~~ validate probabilistic forecasts and can been seen as generalization of the mean absolute error to the probabilistic realm of ensemble forecasts. It is ~~described~~ defined as:

$$CRPS = \frac{1}{N} \sum_{n=1 t=1}^{N} \int_{-\inf}^{+\inf} \left[F_t(x(n)) - H_t(x(n) - x_0)^2 \right] dx \qquad (1)$$

where ~~x(n) is the nth forecast of the~~ $x(n)$ is the forecast at time step t of N number of forecasts and x_0 is the observed value. The CRPS is the continuous extension of RPS where $F(x)$ is the cumulative distribution function (CDF) $F(x) = p(X - x)$ and $H(x - x_0)$ is the Heaviside function, which has the value 0 when $x - x_0 < 0$ and 1 otherwise.

The CRPS compares the cumulative probability distribution of the discharge forecasted by the ensemble forecast system to an observation. It is sensitive to the mean forecast biases as well as the spread of the ensemble. Since the SEAM has 11 ~~ensemble members were randomly drawn from the~~

SEAS ensemble to have the same number of ensemble members as in SEAM. To account for the difference in ensemble size between SEAM members and SYS4 and CLIM has 15 members in the hindcast, the CRPS are not directly comparable. Ferro et al. (2008) showed that for two ensemble distributions with different ensemble sizes, M and m, the unbiased estimate for $CRPS_M$ based on CRPS calculated from the ensemble size m is:

$$CRPS_M = CRPS_m - \frac{M-m}{2Mmn}\sum_{t=1}^{n}\Delta_t \tag{2}$$

where

$$\Delta_t = \frac{1}{m(m-1)}\sum_{i \neq j}|X_{t,i} - X_{t,j}| \tag{3}$$

is Gini's mean difference of ensemble members $[X_{t,1},...,Xt,m]$ at time t. From the CRPS a skill score (CRPSS) can be derived by comparing CRPS of the verified forecast against a reference forecast.

$$SS_{CRPS} = 1 - \frac{CRPS_{fc}}{CRPS_{rf}} \tag{4}$$

In this paper CRPS was Mean relative error was measured as the forecast error against observations normalised against observations. The reliability was assessed through a reliability diagram, where the forecast probability of exceeding the percentiles of its climatology is compared against the observed frequencies Weisheimer and Palmer (2014). All scores were calculated for SYS4, SEAM and CLIM over the hindcast period. CRPSS is used throughout the paper as a measure to calculate the added value of the different forecasts.

3 Results and discussion

3.1 Overall forecast skill

The forecast skill gain provided by SEAM with respect to SYS4 is mostly concentrated to the first 6 six weeks (Figure 3,a) when the forcing data are from the ENS-ER. The difference in CRPSS is 0.6 at week 1one, which then decreases to 0.2 by week 6. six. All river points show a gain in skill up until week 3three, then some points show a benefit of using the SYS4 instead of SEAM. However, in some catchments there is skill up further than 8 eight weeks. The overall better performance of SEAM with respect to SYS4 is partly because of the use of a more recent model version and partly because of the more frequent update of the atmospheric and hydrological initial conditions. It is possible to disentangle the relative contributions between these two factors by only considering a reduced number of starting dates for the SEAM forecast; i.e dates that are the closest to the ~~SEAS4~~ SYS4 starting dates (figure 3,b). This reduced statistic provides a measure of the expected contribution of *only* employing a newer model cycle in the first weeks while both simulations benefits from the same hydrological initialization. In this case the skill gain in CPRS reduces to between 0 and 0.4 (median 0.2) against SYS4 for the first week, reducing to neutral around week ~~4.~~ four. Therefore the most relevant gain comes from the more frequent initializations of the hydrological model.

To put these increments into context we also look at the improvement in skill of the two system SYS4 and SEAM against the CLIM benchmark forecast (Figure 3c-d). The gain from having an improved initial conditions in SEAM is similar in comparison with CLIM (Figure 3c) as ~~compared~~ with SYS4 (Figure 3a) in the first week, but the skill deteriorates quicker and the median CRPSS is negative after 5 weeks. Without the increase in skill due to the advantage in the better initial conditions, SEAM still shows a gain against the CLIM forecast with a CRPSS of 0.4 for the first week, although the spread is quite large (Figure 3d). Also SYS4 shows an increase of skill against the CLIM forecast. Both forecasts are less skillful than CLIM for most river points after week ~~4.~~ four. It can also be noted that SEAM has a higher spread ~~in~~ than SYS4 on longer lead times even though they are forced with the same data from ~~week 7~~ day 47 and onwards. ~~This can when~~ An explanation can be that the ensembles from the two meteorological forecasts are not matched in terms of their relative attributes with regards to their ensemble mean. If two extreme driving forecasts from the two meteorological forecasts are combined it can lead to members that are further away from the ensemble mean than when only one driving forecast is used.

3.2 Geographical variation of forecast skill

The geographical distribution of skill gain provided by the SEAM and SYS4 prediction ~~revels~~ reveals a coherent picture with good scores against the CLIM run over most of Europe (Figure 4 a-b). The gain in the figure is expressed as a difference in the number of weeks into the forecast needed for the CRPSS to drop below zero (i.e. there is no skill in the forecast in comparison with CLIM), which gives an indication of the expected time gain in terms of information provided by the forecast against the reference forecast. Both SYS4 and SEAM are better than CLIM, and SEAM has higher skill than SYS4 for most of Europe. There is a small negative affect over the Alps, southeastern Europe and northern Finland (Figure 4d). The performance of the operational EFAS in these regions is generally poor, which is caused by the difficulty of having good observations of precipitation in high altitude stations and the atmospheric models difficulty in resolving steep orography (Alfieri et al., 2014).“The snow accumulation and snowmelt are further divided into three elevation zones within a grip in LISFLOOD to better account for orographic effects in mountainous regions. However, this increase in sub grid resolution is not likely to be high enough to capture the snow variability during the snow accumulation and snowmelt in mountainous regions. Further, precipitation forecasts have documented biases in steep orography ((Haiden et al., 2014)).

Another interesting aspect to showcase is the relevance of more frequent model version updates is the overall improvement on river discharge for all stations in proximity ~~of the west~~ to the western coasts. This can be attributed to recent developments of the precipitation forecasts, for example a new diagnostic closure introduced in the convection scheme ~~((Bechtold et al., 2014))~~ (Bechtold et al., 2014) and a new parameterization of precipitation formation (Haiden et al., 2014).

3.3 Bias and reliability

The relatively sharp decline in CRPSS can to some extent be explained by the negative bias (too wet forecast) for both SEAM and SYS4 forecast (Figure 5). SEAM has lower bias than SYS4, also when the analysis is confined to the first few weeks (Figure 5b). The slightly better bias in SEAM disappears quickly after the merge (week 7). The bias of the forecast is not spatially consistent, it is generally larger west- and mid-Europe (Figure 6). The figure shows the bias for SEAM (a-c) but the pattern is similar for SYS4. SEAM has generally a smaller bias than SYS4 (Figure 6d). SYS4 has lower bias south of the alps, where it is also performs better than SEAM.

Reliability of a forecast is terms of its usefulness for decision making. A reliable forecast can be trusted to predict the correct probability of certain events, regardless of the accuracy. An unreliable forecast is in practice of no use and can lead to poor decisions Weisheimer and Palmer (2014). Both forecast systems are over-confident (Figure 7, which is similar to a previous study of 2 m temperature and precipitation over Europe with SYS4 Weisheimer and Palmer (2014). The skill of the forecasts from any of the system could be potentially higher by performing a bias correction either of the atmospheric input and/or of the discharge. However in this paper we concentrate on the difference in skills provided by the various configurations and no bias correction has been applied.

3.4 Added value of the seamless forecast

Even though the increase in the overall skill provided by the SEAM in comparison with SYS4 is noticeable, ~~however~~ the justification for its use in an operational context also depends on the actionable time gain in a response situation. More frequent forecast updates could potentially be useful in decision making. As an example we analyze the predicted stream flow for the Rhine river at a station just upstreams Cologne, Germany, during the ~~exceptional~~ European heat wave in the summer of 2003. It was an exceptional meteorological event which combined significant precipitation deficits with record-setting high temperatures (García-Herrera et al., 2010). At its peak ~~,~~ in August, extremely low discharge levels of rivers were reported in large parts of Europe. Economic losses where huge in many primary economic sectors including transportation (Ciais et al., 2005). For several months inland navigation was heavily impaired and the major European transport routes in the Danube and Rhine basins ceased completely (Jonkeren et al., 2008). ~~The navigations on the Rhine is not allowed if the water levels reach a certain upper limit but there is no restrictions on the lower water limit (Meißner et al., 2017).~~

Despite the fact that 2003 conditions were extreme ~~,~~ from the meteorological point of view, the upcoming deficit in precipitation and the high temperatures were well predicted by the ECMWF seasonal systems operational at that time (System-3~~) (Weisheimer et al., 2011)~~; Weisheimer et al. 2011). The good predictability of the event is confirmed by the low discharge prediction provided by SYS4 at the Rhein upstreams of Cologne (figure 8). More then 30 % of the ensemble members forecast extreme low-flow conditions. In fact the observed discharge confirms that the river flow on two separate occasions, event ~~1~~ one on August 17-27 and event ~~2~~ two September 18-28 2003, went below the 3% percentile of its climatological value for the season (figure 8). While most of SYS4 ensemble members mark the extreme condition ~~with 3 to 4 week~~ three to four weeks ahead, there is no information of the recovery period observed between event ~~1 and 2~~ one and two in the forecast starting the first of August. SYS4 predicts a swift recovery back to normal conditions on the forecast issued ~~the~~ 1 ~~of~~ September. A more detailed picture of this intermediate recovery is instead conveyed by the seamless system. Thanks to the more frequent updates, the temporary increase in river flow is correctly picked-up giving a potential advantage of ~~2 to 3~~ two to three weeks for planning actions. SYS4 does indicate the second low flow with a longer lead time than SEAM. However, SYS4 misses the timing of the event.

Even if this was a good forecast for ~~System 4~~SYS4, the information it provides is more informative (anomaly condition) than "actionable" (White et al., 2017). In the above example, a decision maker would have to make a decision based on a forecast that was issued 2.5 weeks earlier, which would inherently make the decision more uncertain if you only had the seasonal forecast. With the seamless system available a decision maker would gain the same early indication of a hazardous event and also have the benefit of frequent updates. In this particular case, the SEAM forecast for the first event was more unstable for some ensemble members, but in general the event was well captured (Figure 8). The SEAM could also correctly capture the recovery with higher water levels between the extreme low flow events. The onset of the second low period was correctly modeled by the SEAM system, whereas the timing of the low flow was missed by SYS4~~did not predict it in with the right timing~~. It should be said that using other less extreme thresholds (<90 and <95 percentiles) even further strengthened the case for using the SEAM.

4 Conclusions

This study compared a set of ~~hydrologic~~ hydrological hindcast experiments over the European domain with two meteorological forcings; ECMWF's seasonal forecasting system (SYS4) and a merged system of ECMWF extended range forecast and seasonal forecast system (SEAM). The latter showed a better overall skill and lower bias over most areas in Europe ~~, and the for~~ with lead times up to ~~7~~ seven weeks. This increase in skill could be attributed to better initial conditions of the hydrological and meteorological model as well as a better atmospheric model version in SEAM.

~~Some~~ In some areas, particularly in the Alps and northern Finland~~where~~ , the seasonal forecast outperformed the merged forecast. However, in these areas the predictability the hydrological model is generally poor which makes these results quite uncertain. Given that the skill in the sub-seasonal range over Europe is in the range of the extended-range ensemble forecast would motivate to use the ENS-ER instead of SYS-4 for hydrometeorological predictions. ~~Still~~

Still, there is an added benefit of using a seamless forecast over the extended range due to the extension of forecast horizon for the early detection of upcoming anomalous conditions. Indeed, as an example this study also highlighted the potential for the use of a sub-seasonal to seasonal ~~for the transport sector in long-term planning. With the higher frequency of and skill ,~~ forecast in the case of an extreme low-flow situation in the River Rhine. The higher frequency and skill of SEAM has the advantage of being a more "actionable" forecast than seasonal forecasts, given that a decision maker would be able to make use of the extra information. Care should be taken when using the forecasts in decision making since the reliability over Europe is "marginally useful" Weisheimer and Palmer (2014). It is therefore important to assess the reliability and skill of SEAM at the location it is to be implemented over the season of interest.

Future work with the seamless forecasting system is to further explore the limits of predictability, reliability and bias to assess the strengths and limitations of the current setup. The assumption that the forecasts can be randomly concatenated would also need to be tested against a system where the forecasts are matched according to their respective climatology. Bias correction of the forecasts might be a necessity, and the advantage of the extended-range and seasonal forecasts from ECMWF is that the availability of hindcasts which are enables just that.

Acknowledgements. This paper was financed through the Framework service contract for operating the EFAS computational centre in support of to the Copernicus Emergency Management Service (EMS)/Early Warning Systems (EWS) 198702. ~~Data from then SEAM and SYS4 experiments are stored in the ECFS repository at ECMWF therefore to access the data an account at ecmwf is required. This can be obtained free of charge for research purposes. However, the whole dataset volume is of the order of 75TB therefore, alternatively, a request to the corresponding author can be submitted and smaller requests can be made down-loadable from a temporary ftp. Finally, these data will be made available through the Copernicus C3S climate data store (CDS), which is under development.~~ The authors would ~~also like to thank~~ like to acknowledge Blazej Krzeminski for setting up the computational framework and Florian Pappenberger for the discussions regarding the seamless forecast system. We would also like to thank Kean Foster, Bastian Klein and Mike Hardeker for useful comments on the discussion paper.

References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, Journal of Hydrology, doi:http://dx.doi.org/10.1016/j.jhydrol.2014.06.035, http://www.sciencedirect.com/science/article/pii/S0022169414004958, 2014.

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrology and Earth System Sciences Discussions, 2017, 1–27, doi:10.5194/hess-2017-610, https://www.hydrol-earth-syst-sci-discuss.net/hess-2017-610/, 2017.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: a global land surface reanalysis data set, Hydrology and Earth System Sciences, 19, 389–407, doi:10.5194/hess-19-389-2015, http://www.hydrol-earth-syst-sci.net/19/389/2015/, 2015.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141–153, doi:10.5194/hess-13-141-2009, http://www.hydrol-earth-syst-sci.net/13/141/2009/, 2009.

Bazile, R., Boucher, M.-A., Perreault, L., and Leconte, R.: Verification of ECMWF System 4 for seasonal hydrological forecasting in a northern climate, Hydrology and Earth System Sciences, 21, 5747–5762, doi:10.5194/hess-21-5747-2017, https://www.hydrol-earth-syst-sci.net/21/5747/2017/, 2017.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.-P., Beljaars, A., and Bormann, N.: Representing equilibrium and nonequilibrium convection in large-scale models, Journal of the Atmospheric Sciences, 71, 734–753, 2014.

Brown, A., Milton, S., Cullen, M., Golding, B., Mitchell, J., and Shelly, A.: Unified modeling and prediction of weather and climate: A 25-year journey, Bulletin of the American Meteorological Society, 93, 1865–1877, 2012.

Brunet, G., Shapiro, M., Hoskins, B., Moncrieff, M., Dole, R., Kiladis, G. N., Kirtman, B., Lorenc, A., Mills, B., Morss, R., et al.: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction, Bulletin of the American Meteorological Society, 91, 1397–1406, 2010.

Burek, P., Van Der Knijff, J. M., and De Roo, A.: LISFLOOD - Distributed Water Balance and Flood Simulation Model - Revised User Manual 2013, EUR - Scientific and Technical Research Reports 978-92-79-33191-6 (print); 978-92-79-33190-9, doi:10.2788/24982, 2013.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., et al.: Europe-wide reduction in primary productivity caused by the heat and drought in 2003, Nature, 437, 529, 2005.

De Roo, A. P. J., Wesseling, C. G., and Van Deursen, W. P. A.: Physically based river basin modelling within a GIS: the LISFLOOD model, Hydrological Processes, 14, 1981–1992, doi:10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F, http://dx.doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO, 2000.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, doi:10.1002/qj.828, http://dx.doi.org/10.1002/qj.828, 2011.

Di Giuseppe, F., Molteni, F., and Tompkins, A. M.: A rainfall calibration methodology for impacts modelling based on spatial mapping, Quarterly Journal of the Royal Meteorological Society, 139, 1389–1401, 2013.

Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., Barbosa, P., Vogt, J., and Pappenberger, F.: Global meteorological drought – Part 2: Seasonal forecasts, Hydrol. Earth Syst. Sci., 18, 2669–2678, doi:10.5194/hess-18-2669-2014, http://www.hydrol-earth-syst-sci.net/18/2669/2014/, 2014.

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

García-Herrera, R., Díaz, J., Trigo, R., Luterbacher, J., and Fischer, E.: A review of the European summer heat wave of 2003, Critical Reviews in Environmental Science and Technology, 40, 267–306, 2010.

Haiden, T., Magnusson, L., Tsonevsky, I., Wetterhall, F., Alfieri, L., Pappenberger, F., de Rosnay, P., Muñoz-Sabater, J., Balsamo, G., Albergel, C., Forbes, R., Hewson, T., Malardel, S., and Richardson, D.: ECMWF forecast performance during the June 2013 flood in Central Europe, Report, Euopean Centre for Medium-Range Weather Forecasts, 2014.

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., Wyser, K., Dutra, E., Baldasano, J. M., Bintanja, R., et al.: EC-Earth: a seamless earth-system prediction approach in action, Bulletin of the American Meteorological Society, 91, 1357–1363, 2010.

Hazeleger, W., Wang, X., Severijns, C., Ştefănescu, S., Bintanja, R., Sterl, A., Wyser, K., Semmler, T., Yang, S., Van den Hurk, B., et al.: EC-Earth V2. 2: description and validation of a new seamless earth system prediction model, Climate dynamics, 39, 2611–2629, 2012.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, http://dx.doi.org/10.1175/1520-0434(2000)0152.0.CO;2, doi:10.1175/1520-0434(2000)0152.0.CO;2, http://journals.ametsoc.org/doi/abs/10.1175/1520-0434(2000)015%3C0559%3ADOTCRP%3E2.0.CO%3B2, 2000.

Hoskins, B.: The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science, Quarterly Journal of the Royal Meteorological Society, 139, 573–584, 2013.

Hurrell, J., Meehl, G. A., Bader, D., Delworth, T. L., Kirtman, B., and Wielicki, B.: A unified modeling approach to climate system prediction, Bulletin of the American Meteorological Society, 90, 1819–1832, 2009.

Jonkeren, O., van Ommeren, J., and Rietveld, P.: Effects of low water levels on the river Rhine on the inland waterway transport sector, in: Economics and Management of Climate Change, pp. 53–64, Springer, 2008.

Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in Central Europe, Hydrology and Earth System Sciences Discussions, 2017, 1–31, doi:10.5194/hess-2017-293, http://www.hydrol-earth-syst-sci-discuss.net/hess-2017-293/, 2017.

Molteni, F. and co authors: The new ECMWF seasonal forecast system (System 4), Tech. rep., ECMWF.

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), Report, ECMWF, 2011.

Ntegeka, V., Salamon, P., Gomes, G., Sint, H., Lorini, V., Thielen del Pozo, J., and Zambrano, H.: EFAS-Meteo: A European daily high-resolution gridded meteorological data set for 1990 - 2011, Report, doi:10.2788/51262, 2013.

Palmer, T. and Webster, P.: Towards a unified approach to climate and weather prediction, in: Proceedings of 1st Demetra Conference on Climate Change, 1993.

Palmer, T. N., Doblas-Reyes, F. J., Weisheimer, A., and Rodwell, M. J.: Toward seamless prediction: Calibration of Climate Change Projections Using Seasonal Forecasts, Bull. Amer. Meteor. Soc., 89, 459–470, 2008.

Pappenberger, F., Wetterhall, F., Dutra, E., Di Giuseppe, F., Bogner, K., Alfieri, L., and Cloke, H. L.: Seamless forecasting of extreme events on a global scale, pp. 3–10, missing, 2013.

Sene, K., Tych, W., and Beven, K.: Exploratory studies into seasonal flow forecasting potential for large lakes, Hydrology and Earth System Sciences, 22, 127–141, doi:10.5194/hess-22-127-2018, https://www.hydrol-earth-syst-sci.net/22/127/2018/, 2018.

Smith, P., Pappenberger, F., Wetterhall, F., Thielen, J., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M., and Baugh, C.: On the operational implementation of the European Flood Awareness System (EFAS), Report 778, European Centre for Medium-Range Weather Forecasting, http://www.ecmwf.int/en/elibrary/16337-operational-implementation-european-flood-awareness-system-efas, 2016.

Thielen, J., Bartholmes, J., Ramos, M. H., and de Roo, A.: The European Flood Alert System - Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13, 125–140, doi:10.5194/hess-13-125-2009, http://www.hydrol-earth-syst-sci.net/13/125/2009/, 2009.

Van Der Knijff, J. M., Younis, J., and De Roo, A. P. J.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, doi:10.1080/13658810802549154, http://dx.doi.org/10.1080/13658810802549154, 2010.

Vitart, F., Buizza, R., Alonso Balmaseda, M., Balsamo, G., Bidlot, J. R., Bonet, A., Fuentes, M., Hofstadler, A., Molteni, F., and Palmer, T. N.: The new VarEPS-monthly forecasting system: A first step towards seamless prediction, Q. J. R. Meteorol. Soc., 134, 1789–1799, 2008.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, Journal of The Royal Society Interface, 11, doi:10.1098/rsif.2013.1162, http://rsif.royalsocietypublishing.org/content/11/96/20131162, 2014.

Weisheimer, A., Doblas-Reyes, F. J., Jung, T., and Palmer, T.: On the predictability of the extreme summer 2003 over Europe, Geophysical Research Letters, 38, 2011.

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

Zajac, Z., Z.-B. M. S. P. B. P. G. A. and Bianchi., A.: Calibration of the lisflood hydrological model for europe - calibration round 2013., 2013.

Table 1. technical details of the forecast and the hindcast used in this paper.

System	T Res	Spatial Res	Horizon	Ens size	Issue frequency	Hindcast set	Hindcast Ens size
ENS-ER	3h/6h	18/36 km[1]	46 days	51	Twice weekly	20 years	11 members
SYS4	6h	80 km	7/13 months	51	Monthly	30 years	15/51 members
SEAM	6h	5 km	6 months	51	Twice weekly	20 years	11 members

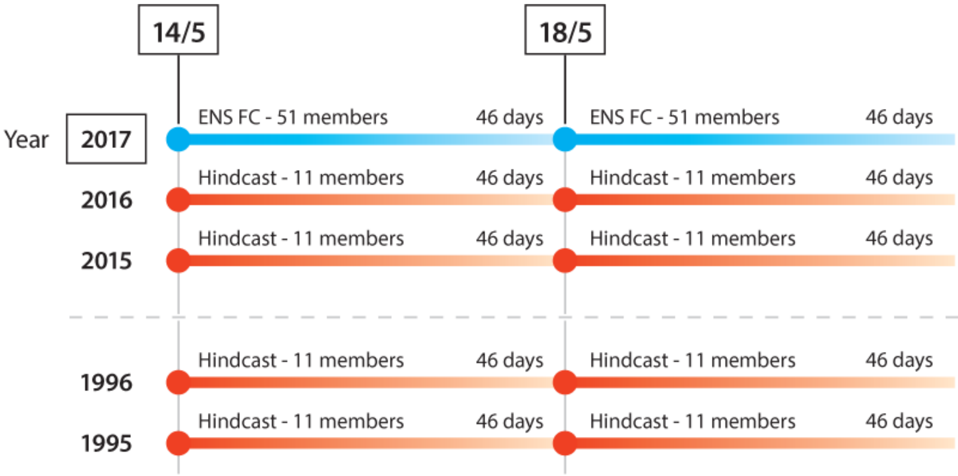


Figure 1. Schematic overview of the operational ECMWF ensemble forecast for the extended range and its associated hindcast. The hindcasts consists of a reduced ensemble forecast (11 members) with the same starting date of year as the current forecast, but run for the previous 20 years.

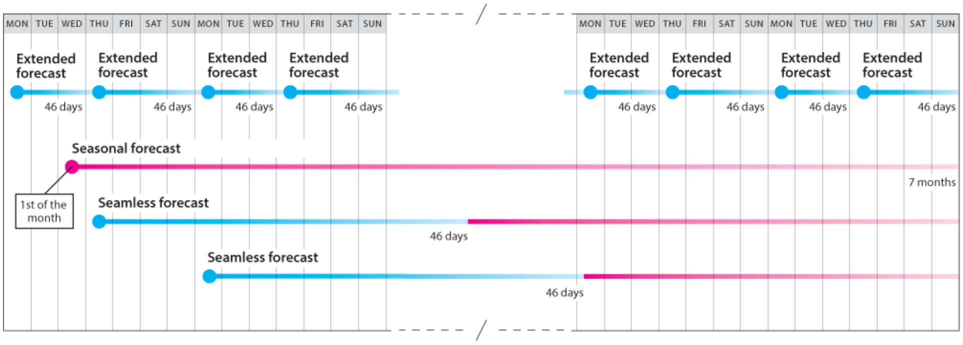


Figure 2. Schematic overview of the seasonal, extended-range forecast and merged systems. The Extended forecast is issued every Monday and Thursday and extends up until 46 days, the seasonal forecasts is issued on the first of each month and extends up until 7 months (13 months in February, May, August and November). The merged forecasts concatenates the latest extended forecast with the latest seasonal forecast.

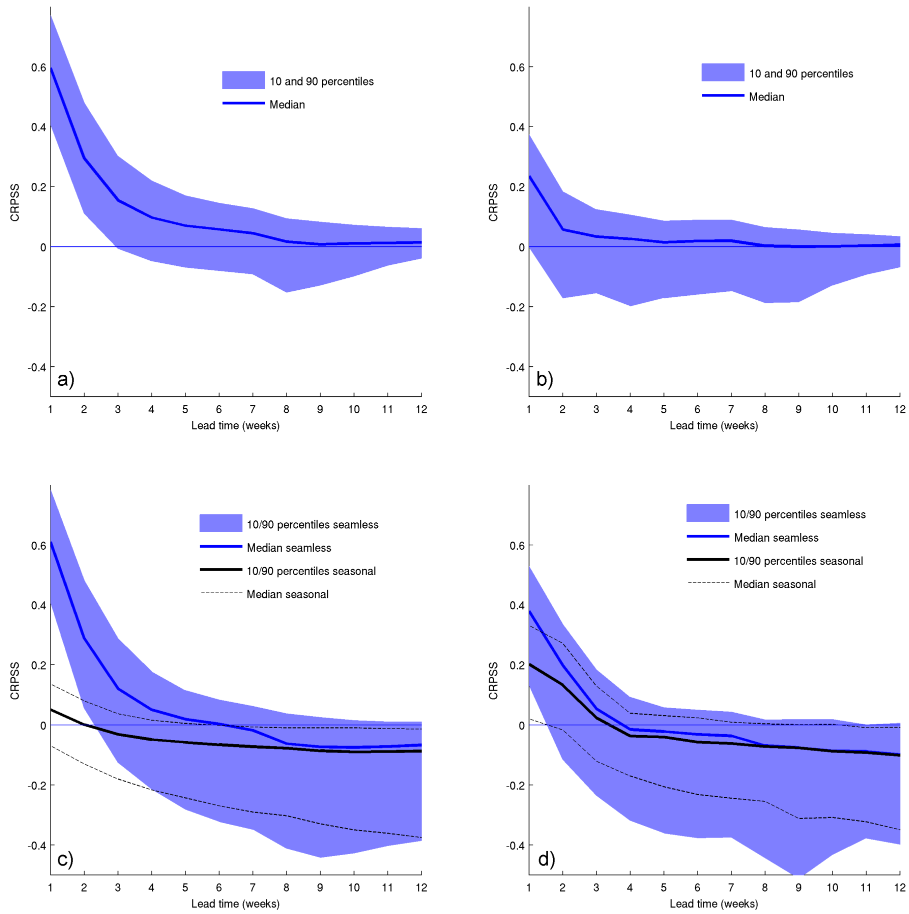


Figure 3. Continuous ranked probability skill score (CRPSS) for a) merged forecast against seasonal forecast for all start dates; b) as in a) but only for the first merged forecast of each month; c) merged forecast against climatology for all lead times in blue and d) as in c) but for the first merged forecast in the month. The shaded blue area denotes the 10-90 percentile of the CRPSS and the blue line the median. The black solid (dotted) lines in figure c and d denote the mean and 10-90th percentile of the CRPSS of the seasonal against the climatological forecast.

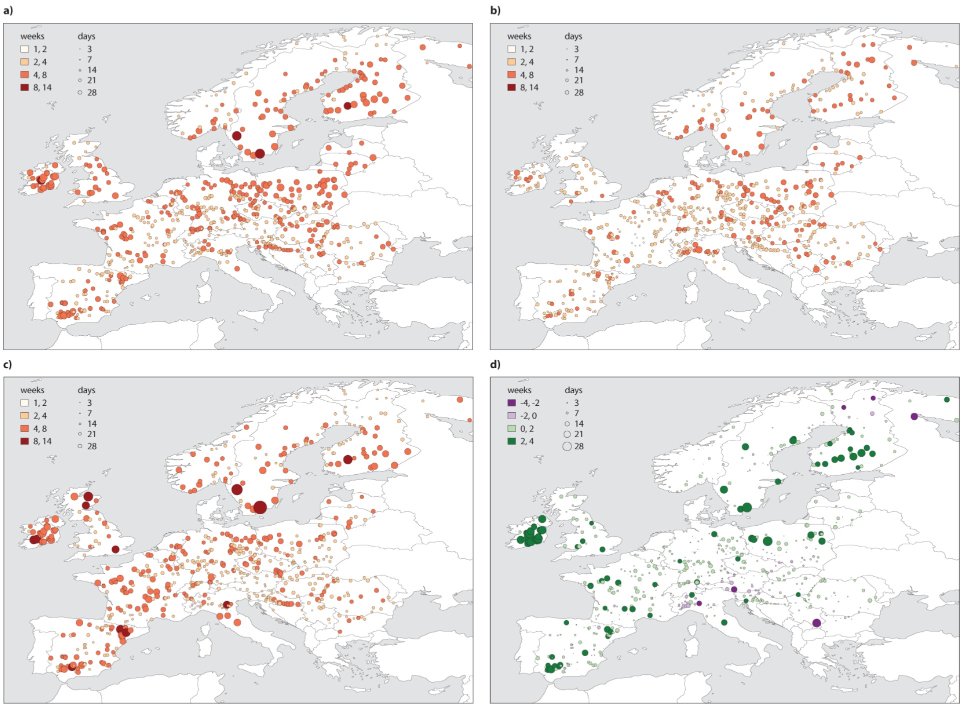


Figure 4. The number of weeks before the CRPSS goes below zero using only the first forecast of the month for a) SEAM against CLIM; b) SYS4 against CLIM c) SEAM against SYS4; and d) difference between SEAM against CLIM and SYS4 against CLIM expressed in weeks. The dimension of the circles is proportional to the number of days while the color scale refers to progressive weeks.

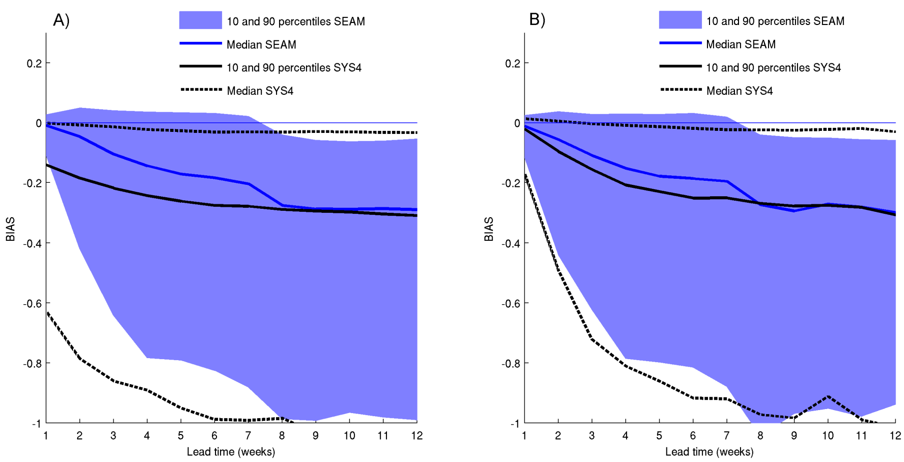


Figure 5. Mean relative error over all outlet points as a functionality of lead time in weeks for a) all starting dates of the forecasts and b) for the starting dates close to the beginning of the months. Negative values denote that the forecast is too wet in comparison with the CLIM run. The SEAM (SYS4) forecast is in blue (black) where the solid line denotes the median and the filled area (area between dotted lines) denote the 10th to 90th percentile.

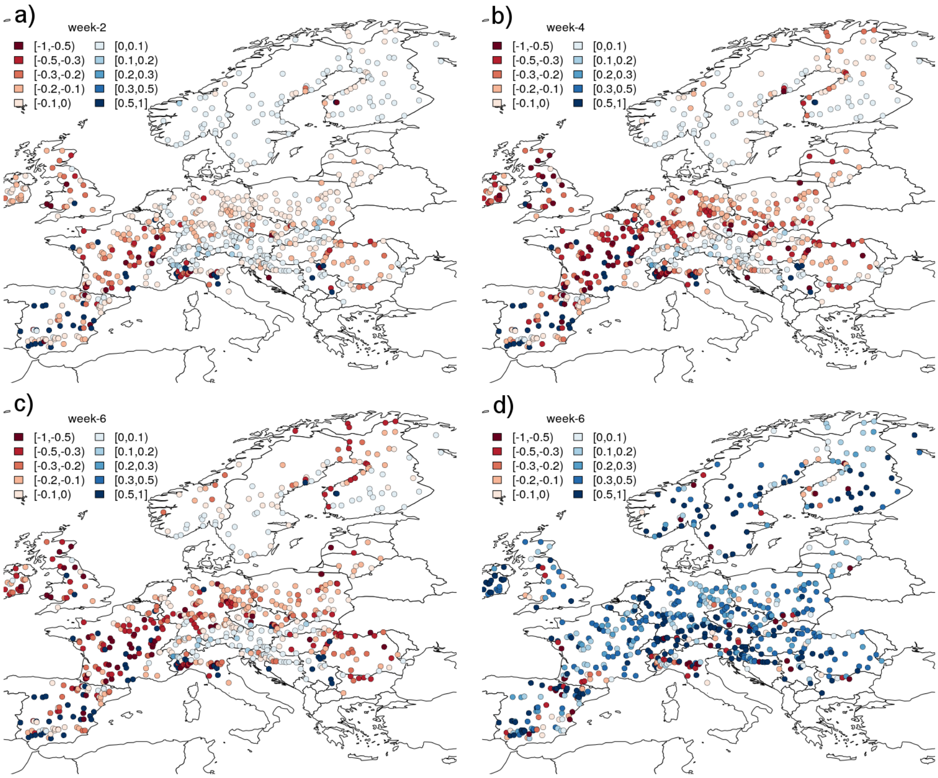


Figure 6. Mean relative error for each of the outlet points for the SEAM forecast over the outlet points for a) week 2, b) week 4 and c) week 6. Red indicates where the forecast is too wet, and blue where it is too dry. Figure d) shows the difference in absolute error between SEAM and SYS, where blue (red) denotes points where SEAM has a smaller (larger) MAE than SYS4.

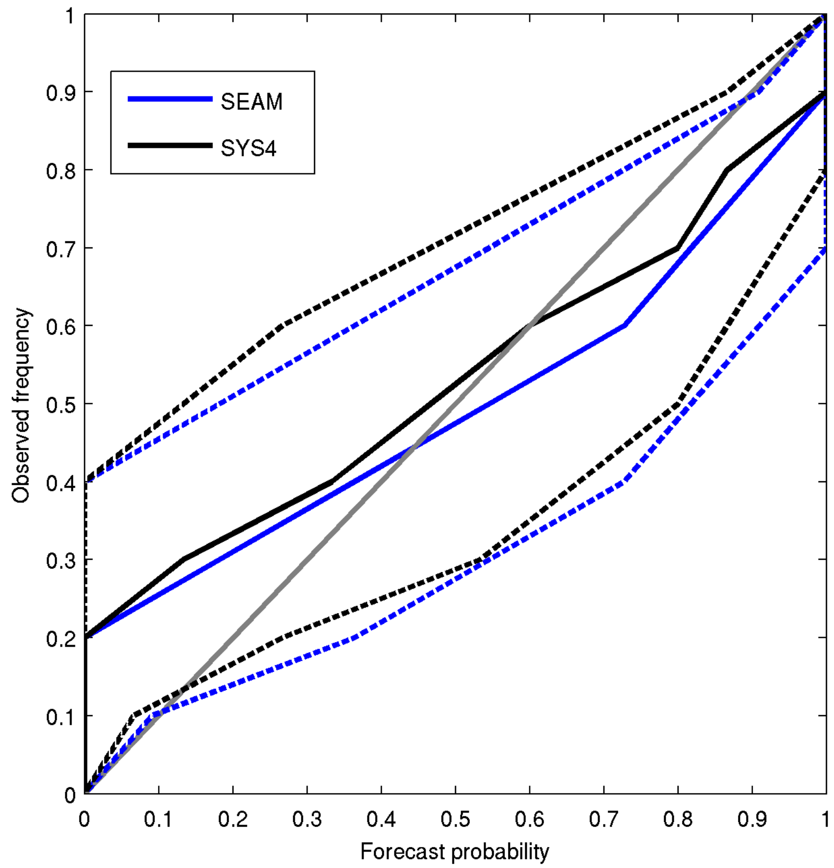


Figure 7. Reliability diagram for SEAM (blue) and SYS4 (black) for week 5 for all outlet points. The solid lines indicate the median reliability and the dotted lines the 25th and 75th percentiles.

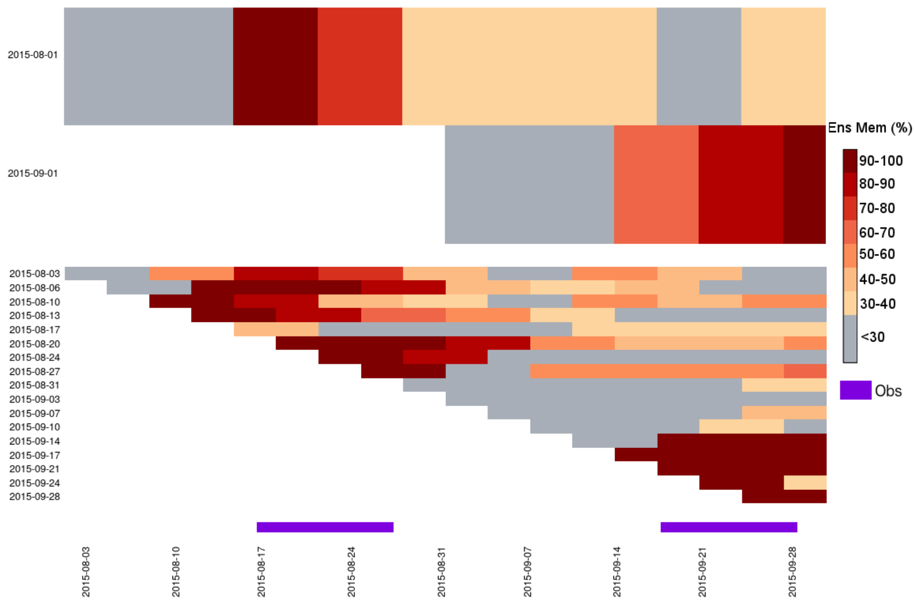


Figure 8. Percentage of ensemble members predicting low flow anomaly (< 97%) on the Rhine river north of Cologne for summer 2003. The two starting dates in August and September from SYS4 are compared to the 17 starting dates of the seamless forecasting system. In two separate events the discharge was recorded below the 97 % percentile, event 1 on 17-27 of August and event 2 on 18-28 of September 2003.