## 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