Harrigan et al. (2017) first response to reviewers - Reviewer 1: RC1 (Anonymous)

All reviewer 1 comments are labelled consecutively, for example, comment 1 is R#1-1, with our responses to reviewers given in blue text.

General Comments:

R#1-1. Overall the paper is well written and makes a positive contribution to the scientific literature within this field. It is well balanced, set out clearly and has a good range of figures. The authors need to address whether they are referring to 'forecasts' or 'projections'. Without conditioning ESP results according to forecast large scale climatic influences i.e. NAO then the results should be termed 'projections' not 'forecasts'. I recommend than with minor revisions the paper should be accepted.

We thank the reviewer for their positive and constructive review. We have made the majority of your suggestions and clarify any points raised below. We address your comment about referring to ESP as a forecast below.

Specific Comments:

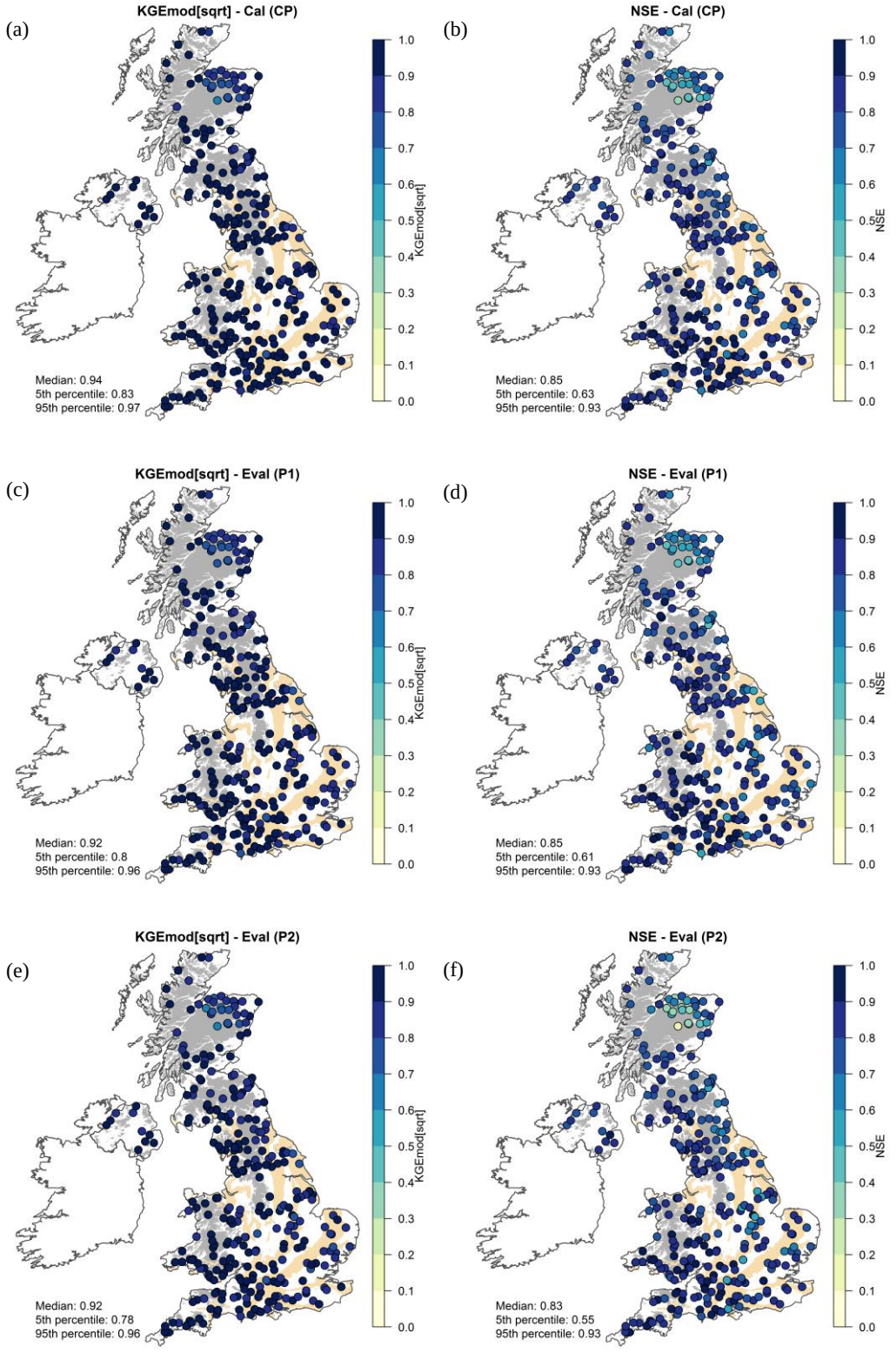
R#1-2. 1. The paper on many occasions refers to 'ESP forecasts', however as this method is not driven by a meteorological forecast it would be better to refer to these as 'ESP Projections'.

Whilst it is true that ESP does not contain any information about future atmosphere dynamics, it is now standard practice to describe its application in terms of a forecast (e.g., wood et al. (2016), as well as papers within this special issue: e.g., Beckers et al. (2016), Crochemore et al. (2017), and Arnal et al. (2017)). We would like to keep our terminology consistent with these papers but could change it if deemed necessary by the editor.

R#1-3. 2. Page 5 lines 11-17: There needs to be greater in depth discussion as to the results presented in Table 2 in the context of other studies. Are the calibration results better than other models/studies?

The main focus of the paper is not on the hydrological modelling component, it is instead to show that the GR4J model used here could reasonably simulate river flow observations in a wide range of catchments across the UK and could be deemed a viable model for catchment-scale ESP forecasting. The particular focus was on calibration and evaluation of medium range flows (hence why the modified Kling-Gupta efficiency applied to root transformed flows KGE_{mod}[sqrt] was used (i.e. Pg5; L3), and not low (e.g. using log transformed flows) or high flow (e.g. using Nash-Sutcliffe Efficacy (NSE)) metrics, as the method aims to provide ESP forecasts across the full range of the flow regime.

However, we acknowledge that it would be useful to know how our modelling results compare to other models/studies. The most universally used metric for hydrological model calibration/evaluation is the NSE. We have therefore also calculated the NSE for all 314 catchments and will provide a summary of results in **supplementary Figure S1** (see below) and will add individual catchment NSE scores for the calibration and evaluation periods, along with KGE_{mod}[sqrt], in **supplementary Table S1** so that others can make more detailed comparisons.



Supplementary Figure 1: Spatial distribution of GR4J model performance for 314 catchments over the calibration (Cal CP [WY1983-2014], top row), and two evaluation periods (Eval P1 [WY1983-1998], middle row and Eval P2 [WY1999-2014], bottom row) for the modified Kling-Gupta efficiency applied to root squared transformed flows (KGEmod[sqrt]) and Nash-Sutcliffe efficiency (NSE) model performance metrics. UK-wide Summary statistics are given in the bottom left for the median and 5th and 95th percentiles.

We propose therefore to insert the following section of text to reflect this review comment on Pg5; L12: "Overall, GR4J performs well against streamflow observations and parameter sets remain stable across P1 and P2 with comparable performance to Crochemore et al. (2017) using GR6J for 16 catchments across France. The NSE was also calculated as it is the most university used metric and so allows comparison to a wider set of studies. Spatial maps and summary statistics of KGE_{mod}[sqrt] and NSE are provided in supplementary Figure S1 and, notwithstanding differences in study design, results for GR4J are on par with other large-sample catchment modelling studies in the UK (e.g., Crooks et al. (2009) using the Probability Distributed Model (PDM; Moore, 2007) for 120 catchments)".

R#1-4. 3. Page 6 Section 3.4: a. Please can the authors clarify what river flow metric are the skill scores being applied to? Is it the skill in comparing the mean daily river flow on a future day 1 day/3day/1 week/2 week etc ahead? Or is it the volume of discharge over the next day/3 days, 1 week/2 weeks,…12 months? b. Did the authors consider using RoC scores to assess skill? Please indicate in the discussion why these were not used.

a.) We thank the reviewer for pointing out that this needs more clarification in the manuscript (which was also queried by 'R#3-3'). The streamflow time-series the evaluation metrics are calculated on is equivalent to the volume of water which flowed from the first (forecast initialisation date) to the last day of the forecast. For simplification, it is expressed in the manuscript in equivalent average daily streamflow (evaluation results are identical for both). We will insert the following text after Pg 5; L25 for clarification: "Note that lead time in this paper refers to the aggregation of mean streamflow over the period from the forecast initialisation date to n days/months ahead in time. So a January ESP forecast with 1-month lead time is the mean streamflow from 1 January to the end of January and a January forecast with 2-month lead time is the mean streamflow from 1 January to end of February".

b.) The choice of score to evaluate forecast skill is always a difficult subject; in Wilks (2011), the forecast verification chapter on the plethora of available scores/metrics is nearly 100 pages long. The main aim of our work was to investigate the overall performance of the ESP method, and quite rightly pointed out in the 'R#3-7' comment it is an ensemble forecasting method so focus should be on probabilistic scores – we've used one of the most common metrics, the Continuous Ranked Probability Score (CRPS, and skill score) which is a proper score and has the advantage of defaulting to the Mean Absolute Error (MAE) for a deterministic forecast, so is easy to interpret. The ROC diagram and the area under the ROC curve are indeed another way to evaluate the probabilistic forecast performance, but we chose CRPSS for the above reasons.

We have undertaken additional assessment on the use of different forecast evaluation metrics based on suggestions from Reviewer #3 and have taken on board their recommendation to concentrate on the CRPSS instead of the MSESS in the revised manuscript (please see our responses to R#3).

Technical Corrections:

R#1-5. Page 2 line 10: The Environment Agency implemented operational ESP groundwater level projections in March 2012.

Think will be inserted in Pg2; L13: i.e., "…and also feeds into the Environment Agency's monthly 'Water Situation Reports for England' (operational for groundwater levels in March 2012)".

R#1-6. Page 3 line 28: 'NHMP 2017' is the wrong font size

Will change in the revised manuscript.

R#1-7. Page 4 line 9: 'hydro climatic regions' – how have these been defined and by whom? please include the reference for their designation.

The hydroclimatic regions used in the manuscript were defined based on merging contiguous UK hydrometric areas, which are integral river catchments having topographical similarity with outlets to the sea/estuaries (NRFA, 2014), into regions that reflect broad hydrological and climatological patterns in the UK. The approach was based on expert judgment and guided by the Met Office UK regional precipitation regions (HadUKP: https://www.metoffice.gov.uk/hadobs/hadukp/). For example, the division between North-west England & North Wales (NWENW) and South-west England & South Wales (SWESW).

Note that these UK Hydroclimate regions were designated to facilitate the analysis and interpretation of the results, and in particular to investigate if any ESP skill patterns emerged in contrasting hydroclimatic regions. They have, however, no impact on the individual forecast performance. We will edit the revised manuscript on Pg4; L9 for clarity by inserting the following text: "The nine UK Hydroclimate regions were derived by merging contiguous UK hydrometric areas (NRFA, 2014) that reflect broad hydrological and climatological similarity across the UK and are used for aiding interpretation of results".

The UK Hydroclimate Region shapefile, together with metadata, is openly available from the authors or NRFA (nrfa@ceh.ac.uk), and we also highlight this under Sect. 7 – Data availability.

National River Flow Archive: Integrated Hydrological Units of the United Kingdom: Hydrometric Areas with Coastline, NERC Environmental Information Data Centre, Available from: https://doi.org/10.5285/1957166d-7523-44f4-b279-aa5314163237, 2014.

R#1-8. Page 4 line 13: There are no major sandstone aquifers in Southern England.

We thank the reviewer spotting this. We will remove reference to sandstone.

R#1-9. Page 4 line 16:' highly productive' – please can you provide an explanation to this term

Highly productive refers to highly permeable aquifers (e.g. Chalk). We agree that this does not fit well here as we are referring to a 'Chalk river', and not specifically the aquifer underneath the catchment so will remove 'highly productive' and change the sentence "in catchments with productive aquifers" in P11; L21 to "in catchments with highly permeable aquifers".

When we refer to a catchment with a large groundwater influence on streamflow, we say the catchment is 'slow responding'.

R#1-10. Page 5 line 7: need to define a UK water year (starting 1st October in year in question)

This was mentioned on Pg4; L3, but we will modify to make it more clear: "Q was retrieved from the NRFA over the longest possible period of observed Q across the 314 stations, 32 water years from 1983 to 2014 (water year from 1 October to 30 September designated by the calendar year in which it ends)".

R#1-11. Page 8 lines 14-15, Page 10 lines 28-29 Page 13 lines 9 and 10: There is generally little variation in monthly rainfall across the year – spring and summer are not necessarily significantly drier. It's the greater evaporative demands in the spring and summer which drives the transition referred to.

We thank the reviewer for this comment as quite rightly the transition between these two half year periods is not significant in terms of precipitation, but the increased evaporative demand. This is summarised better in terms of Soil Moisture Deficits (SMDs). So we will change each of these instances to "April, which in the UK is a transition month between winter months with lowest soil moisture deficits (SMDs) and summer months with highest SMDs".

R#1-12. Page 11 line 8: The location of the Mole at Kinnersley Manor will not be known by most readers .It would be better to include the location of all sites mentioned in the text on Figure 1 rather than the insert to Figure 2 which does not include the Mole at Kinnersley Manor.

This is a good suggestion and we will label the 5 catchments mentioned in Figure 2, along with the Mole at Kinnerley Manor, in Figure 1

R#1-13. Figure 1: Include names of sites referred to in the text and Figure 2.

This will be done as per R#1-12, thanks.

R#1-14. Figure 3: Consider a non linear x axis scale to allow readers to view sub monthly skill results – this is not possible with a linear scale.

We believe the linear scale shows the high rate of skill decay and so would rather keep the linear scale. However, we agree that sub-monthly results are too difficult to see. We have therefore redrawn Figure 3 to include results for short (1- and 3-days) and extended (1- and 2-weeks) lead times (below). Note: this figure is now based only on CRPSS based on R#3 comments on most appropriate choice of skill score.

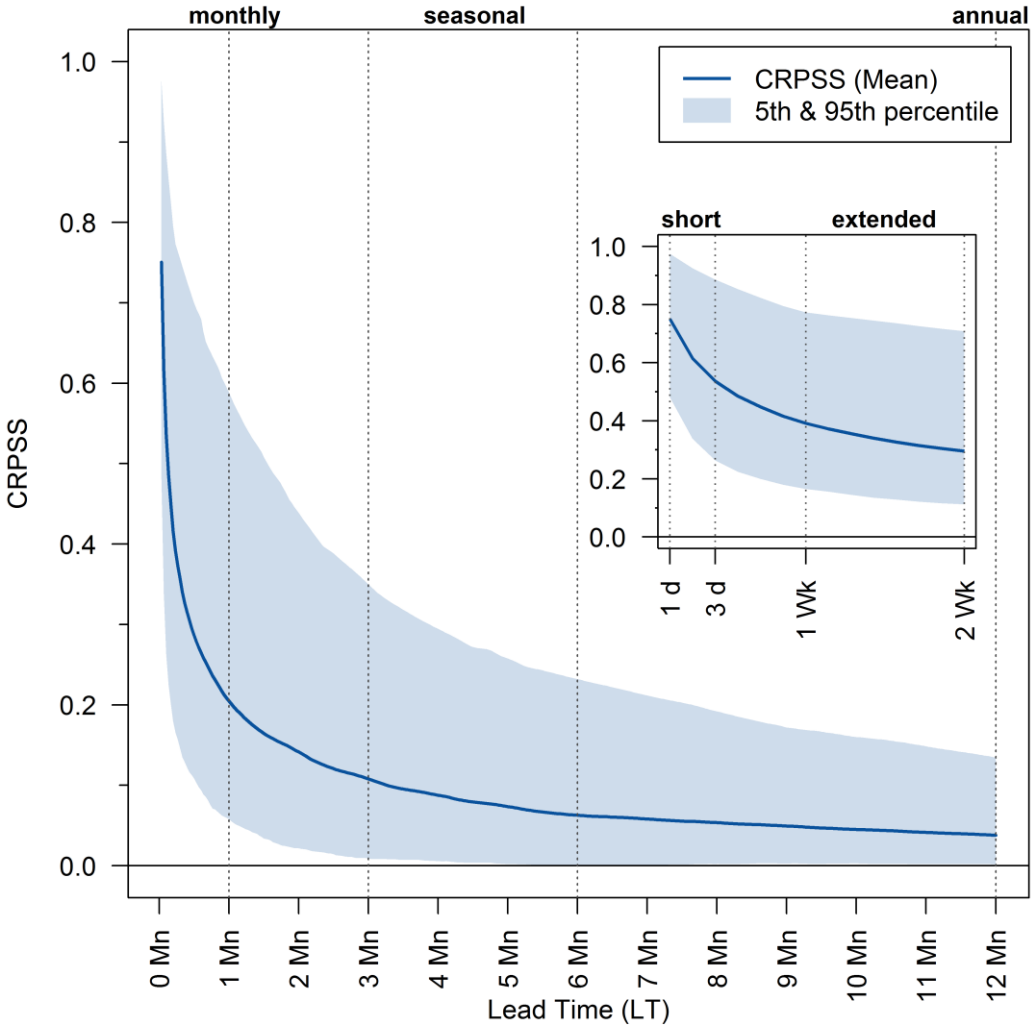


Figure 3: UK-wide mean ESP CRPSS values across all 314 catchments and 12 forecast initialisation months for all 365 lead times (LTs) with short and extended lead times also shown inset for readability. The range of skill scores across catchments at each LT is shown by the semi-transparent 5th and 95th percentile band. Vertical lines represent eight commonly used operational forecasting LTs from short (days) to annual (12-months).

R#1-15. Figure 8: axis labels are absent on all x and y axis – is this because they are dimensionless, if not please can these be included on the figure?

Figure 8 has now been modified based on reviewer comment R#2-5. X1 (mm) and X3 (mm) are now combined are now combined as catchment storage capacity (X1 + X3 in mm) but the log is taken due to it being heavily skewed (as was the case for these variables using in the original manuscript). Therefore the units are 'log mm'. BFI and CRPSS are dimensionless '[-]'. As per your suggestion, axis labels have now been included and will be updated in the revised manuscript accordingly.

Thank you again for your constructive comments,

Shaun.

References

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B. and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrol. Earth Syst. Sci. Discuss., 2017, 1–27, doi:10.5194/hess-2017-610, 2017.

Beckers, J. V. L., Weerts, A. H., Tijdeman, E. and Welles, E.: ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction, Hydrol. Earth Syst. Sci., 20(8), 3277–3287, doi:10.5194/hess-20-3277-2016, 2016.

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

Crooks, S. M., Kay, A. L. and Reynard, N. S.: Regionalised Impacts of Climate Change on Flood Flows: Hydrological Models, Catchments and Calibration, Centre for Ecology & Hydrology, Environment Agency, Defra, London., 2009.

National River Flow Archive: Integrated Hydrological Units of the United Kingdom: Hydrometric Areas with Coastline, NERC Environmental Information Data Centre, Available from: https://doi.org/10.5285/1957166d-7523-44f4-b279-aa5314163237, 2014.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J. and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, J. Hydrometeor., 17(2), 651–668, doi:10.1175/JHM-D-14-0213.1, 2016.