Response to Reviewer #2

I thank Ms Franz for her feedback and suggestions. The manuscript is substantially improved by the changes described below (her original concerns in bold).

"The author misses the possibility that the good performance at the downstream points may be due to the scale of the forecast basin and the limitations of modelling small watersheds… "

The other reviewer raised a similar issue and so the text has been changed to

Despite the large range of error standard deviations from one location to another, the CP indicates that the skill of forecasts is relatively even across the basin. There is a larger difference in 1- and 5-day ahead CP for the upstream locations than there is for the downstream locations between Kratie and Neak Luong, which may be the attributed to the greater uncertainties in initial conditions, recent and future precipitation and other meteorological influences at the smaller scale watersheds found upstream. Indeed, the lowest performing forecasts (5-days ahead at Chiang Saen) rely almost exclusively on the signal contained in observed upstream flows due to the lack of access to rainfall observations in China. Downstream, where hydraulic routing effects have a greater influence than local precipitation, there is nearly no loss of skill with leadtime. The exception is the two furthest downstream forecast points, where low flow forecasts have relatively high error when the river height is affected by the ocean (e.g. observe the poor performance of Tan Chau forecasts in June-July, relative to those in September-October in Figure 2).

"Following on the previous point, I do not entirely agree with the statement on Page 14445, lines 1-3 that locations with a small range of flow are easier to forecast than locations with a large range"

I agree with the reviewer that observed variance is not the only factor affecting skill. It is one of several factors. However, it is a valid measure of the relative difficulty of the forecasting situation. As such this text and reference were added

While the error standard deviation is a highly relevant evaluation measure for an individual user at a single location, this measure is often highly influenced by the hydrological characteristics of the river and is less influenced by the quality of the forecasts. For example, the difference between maximum and minimum height for Luang Prabang during 2000-2012 is 18.2 meters whereas Tan Chau did not vary by more than 5.0 meters. Murphy (1993) lists the unconditional variance of the observations (“Uncertainty”) as one of ten aspects of forecast quality - highly variable observations are intrinsically more challenging to forecast (in absolute terms) than observations with low variability.

Murphy, A. H.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather and Forecasting, 8, 281-293, 1993.

The original text then moves on to normalized forecast scores that factor out the observed variance.

Page 14451: The last paragraph reads like a statement out of a consulting report submitted to the RFMMC. I suggest making this more general.

The final paragraph has been changed to

These analyses would not be possible without the existence of archived forecasts. Operational agencies are strongly encouraged to systematically preserve historical operational forecasts, as well as observations, in a consistent machine-readable format to facilitate easy processing. If possible, such forecast databases should include official products as well as original model inputs and outputs. Adoption of a culture of continual forecast evaluation helps agencies in demonstrating the value of their forecasts to users and assessing the potential benefits of innovations in their forecasting systems.

Page 14437: Refer to Figure 1 at the beginning of the discussion of Study Locations to make the section more understandable

Accepted as suggested

Page 14439, line 26: In general, the meaning of the "as-is forecasts" and "original forecasts" was not immediately clear, and a better explanation should be provided. The sentence on Line 27 states, that "the latter may contain raw model output and not as-issued forecasts". This refers to the "*isis.xls" file. My understanding from later sections is that the "*Original.xls" file should be the one that contains the raw model output. Following on that, on Page 14440, Line 1, what is a "normally-named file"?

This text has been changed to

Operationally, a new spreadsheet is saved for each day's forecasts, normally named "F" with a suffix of the issue day, month and year (e.g. F21Aug09.xls). File names may have slightly different suffixes (e.g. F21Aug09_Original.xls, F21Aug09_Isis.xls). The latter may contain raw model output and not official forecasts (i.e. forecaster-approved final values that are issued to the public). The suffix "Original" was allowed in the 0.65% of cases that a normal-named file (i.e. with no suffix) did not exist for a given date. 3,531 spreadsheets were identified as potentially containing official forecasts.

Page 14442, line 6: The quality score “proposed” by Plate et al. (2008), seems to be the same presented on page 14445 and attributed to Kitanidis and Bras (1980). Perhaps the word “proposed” is inappropriate here. If they indeed are the same, the same name should be used in both sections.

The other reviewer had similar concerns and so the text was changed to

Plate et al. presented a “Quality Index”, which is similar to NS but uses persistence instead of long-term average water level as a baseline and has a reverse orientation (i.e. 0 is perfect, 1 is no-skill). The formula for this index is the same as the Coefficient of Prediction (CP, described in the next section) except the orientation is reversed. This is a more difficult baseline to outperform and Quality scores at Pakse were 0.47 for 1 day ahead degrading to 0.74 for 5 days ahead (CP of 0.53 and 0.26, respectively) .

Page 14447, line 20: An explanation about how the persistence with trend forecasts are created is needed. How many previous time steps is the linear trend based on?

The text now includes

This study also uses a baseline of persistence extrapolated using the trend of the two observations prior to forecast issuance:

$$\acute{f}_i(\text{loc}, \text{lead}) = \text{o}_i(\text{loc}) + \text{lead} * [\text{o}_i(\text{loc}) - \text{o}_{i-1}(\text{loc})]$$