1. The reviewer`s main concern is the innovation and the scientific importance of this work. In the reviewer`s opinion, the research question of this work (i.e. evaluation of operational flood forecast performance) is bit too general and therefore hard to focus on a specific point. I feel the author need to better justify in what sense his/her study is innovative from the other studies on this topic as well. Maybe to put in an inappropriate way, the article in this current form looks more like an evaluation report for Mekong River Commission rather than a scientifically sound paper. Therefore, the reviewer recommends the author to specify the innovation of this study by highlighting its contribution to the science in the introduction and conclusion.

The existing introduction discussed the value of performance evaluation (identifying priorities for improvement, demonstrating value of investments in system upgrades, communicating forecast uncertainty to users, and determining if research methods have comparable skill to existing techniques). Further, the introduction mentions the call by others to develop "hydrologic forecasting science", of which performance evaluation of operational systems is a component. The introduction also mentioned that a novel aspect of this study is that it is the first published evaluation of the entire history of forecasts at the RFMMC (either in the gray or scientific literature). The other contribution of this study is the creation of a database of forecasts for other researchers to use. The introduction was modified to include this statement: "Finally, the archive of forecasts created by this study should facilitate side-by-side comparisons of novel techniques and existing operational methods. . Published scientific studies of operational hydrologic forecasting system performance have been rare, and this article is an attempt to highlight the importance of such evaluations and to foster discussion between the operations and research communities." Also, the conclusion now contains "Historical forecasts should be conveniently accessible and available to users and, as such, the archive of forecasts developed by this study should be available on request from the Mekong River Commission."

2. The author has provided quite positive conclusion of the evaluation (line 492 to 496). In the review`s opinion, this conclusion might be a bit too optimistic. Indeed, the error looks low. How good a flood forecasting system is might also depend on the scale of the river, flood frequency, the topography feature of floodplain, shape of the valley and user requirements (e.g. different vulnerability depending on local land use near the river reach, GDP per capita etc). Therefore, it is difficult to draw positive conclusion by only looking at the skill scores without considering the above mentioned aspects. The conclusions need more justifications and the appropriateness of the flood forecasting performance should be (at least qualitatively) discussed by taking account the characteristics the study area.

This paragraph was added to the end of the introduction

"There are many dimensions to forecast quality and this study only focused on aspects of accuracy at specific streamgauges of interest. In addition to accuracy, forecasting systems can be evaluated with respect to

- production (e.g. is the forecast process reproducible, documented, and cost effective?)
- credibility (e.g. are the forecasts perceived as honest, impartial and unprejudiced?)
- transmission (e.g. are the forecasts timely, accessible, and available in a consistent format?)
- messaging (e.g. are the forecasts easy to understand, relevant and specific to user vulnerabilities?)

For example, Smith (2009) proposed a holistic framework of performance indicators and benchmarks for the RFMMC, ranging from forecast accuracy to the time of release of the forecasts and the number of visits to the RFMMC website to satisfaction ratings from customers. Forecast agencies should strive to monitor and improve all aspects of forecast quality (not just forecast accuracy) to ensure that the forecasts are fit for the purposes of users' needs."

3. In addition to Comment 2, the evaluation is done by using the average performance measure of the operational forecasting (e.g. standard deviation of 2.5 m upstream of Phnom Penh, Line 383 to 393). How useful are those generalized numbers? There should be more discussion regarding this point.

The generalized numbers were provided for several purposes – 1. The RFMMC already produces tables of such numbers to give their customers an idea of expected forecast accuracy on any given day. Those

numbers calculated on a short record and this study provided a much longer record. 2. Nearly all research studies publish a high level skill score (more often correlation or Nash Sutcliffe than standard deviation of error, but the three are similar) to show how good their systems are, and so we provide a number for them to allow comparisons between their research methods and the operational techniques. Indeed, we have already been approached by a researcher comparing our error standard deviations to those of his techniques. However, if the reviewer's criticism is that these evaluations are too generalized to be useful because they lump together so many forecasts that may not be of interest to the user, then this is the purpose of the complementary analysis on lines 470-495. That section focuses down to the 0.1% of forecasts around the time when the river transitions from no-flood to flood, which is the primary concern of the RFMMC's customers.

4. The reviewer doubt the use of ISIS model (1D hydraulic model) from Stung Treng to the ocean, where the area is characterized by flat river delta and floodplain, meaning that the flood pattern and local hydraulics are hard to fulfill the assumptions of de Saint Venant equations. More justification is needed on adopting ISIS model in the Mekong flood forecasting system.

We are unsure if the reviewer is concerned about if model exists as we have described it (http://www.halcrow.com/isis/documents/case_studies/isis_professional_case_study_vietnam_thailand_cambodia_laos.pdf) or that is actually run operationally (the article says "ISIS is more computationally intensive than URBS and therefore the latter is run routinely whereas ISIS is run for retrospective analyses and as demand arises"). If the concern is that it is a scientific misapplication of the ISIS model, invalidating the assumptions of the St Venant equations, then we hope that this would have been addressed in the review of other studies using this model. For example, here is an article using the Mekong River Commission's ISIS model, published in HESS. http://www.hydrol-earth-syst-sci.net/16/4637/2012/hess-16-4637-2012.pdf. We have changed our manuscript to include this reference.

5. In flat delta and floodplains like Mekong delta, small increase of water level might potentially lead to massive flood inundation extent. Those forecasted water levels might be misleading in this circumstance. Therefore, in order to conduct a comprehensive evaluation, the performance measures mentioned in the paper might not be enough. I understand flood extent evaluation might be hampered by the lack of flood extent data and the difficulties of implementing 2D flood modelling in the operational forecast. But those points might worth to be mention in the recommendations.

We have added paragraph in the conclusion that emphasizes the narrow focus of our study while recognizing that true forecast system performance is to be evaluated holistically (i.e. are the forecasts actually addressing the users' needs?)

"There are many dimensions to forecast quality and this study only focused on aspects of accuracy at specific streamgauges of interest. In addition to accuracy, forecasting systems can be evaluated with respect to

production (e.g. is the forecast process reproducible, documented, and cost effective?)
credibility (e.g. are the forecasts perceived as honest, impartial and unprejudiced?)
transmission (e.g. are the forecasts timely, accessible, and available in a consistent format?)
messaging (e.g. are the forecasts easy to understand, relevant and specific to user vulnerabilities?)

For example, Smith (2009) proposed a holistic framework of performance indicators and benchmarks for the RFMMC, ranging from forecast accuracy to the time of release of the forecasts and the number of visits to the RFMMC website to satisfaction ratings from customers. Forecast agencies should strive to monitor and improve all aspects of forecast quality (not just forecast accuracy) to ensure that the forecasts are fit for the purposes of users' needs."

6. Satellite data is used in the flood forecasting to supplement the gauge data (line 145 to 150). The satellite data is usually associated with low accuracy and low resolution, depending on the cost. How the bias was removed from the satellite data? What is the uncertainty associated to those satellite data?

The manuscript now reads "The RFMMC has developed statistical (regression-based) methods for removing bias from the satellite-based products. And "RFMMC uses several remotely sensed products but the satellite-

based rainfall estimates commonly differ from the in situ measurements and each other by 20-60% on seasonal timescales (or over 200% in extreme cases)."

Minor comments:

1. Line 64: Sentence might be grammatically incorrect. – changed to "Finally, the performance of the forecasts is measured and the implications are discussed." Is the concern is about number agreement, "is" is connected to "performance" not "forecasts".

2. Line 187 to 191: I doubt those technique details in terms of data-preprocessing (in Excel or scripts) are really necessary. Considering that this is a new dataset, we had to indicate how it was collected. The mention of scripts was to show that both automated and manual processes were involved with collecting the data (i.e. it was gathered automatically, but then visualized and reviewed by a human to make sure that the automatic processes were successful. However, the mention of Excel was removed and this sentence now reads "The Bulletin's tables and graphics are created using spreadsheet templates." (to indicate the base source material).

3. Line 232 to 233: Maybe I did not fully understand, it is not clear to me why 32% of forecasts were excluded. In the previous sentence the article mentions 73 out of 353,547 forecasts as outliers needing further manual inspection. The 32% refers to those 73 forecasts (i.e. 23/73 = 32%). This and the next sentence were changed to emphasize this point: "In 23 (32%) of the outlier cases, the Bulletins contained forecasts for a date other than what was indicated by the filename and therefore were excluded. 12% of outlier cases resulted from a keying error (e.g. 9.3 meant to be 6.3)."

4. The title of Section 3 (i.e. Forecast Methods) and Section 6 (i.e. Methods) might be bit misleading. Maybe revise 'Methods' of Section 6 to 'Evaluation Methods'. Changed as suggested to "Performance evaluation methods".
5. Line 448: Why was 70th percentile used? Was there any special reason? Why not 95th or 90th percentile were used?

   The first half of this paragraph gave the justification "However, the existing measure is an established performance indicator at RFMMC and users are familiar with it. Adjusting the benchmarks so that forecasts are typically 50% satisfactory (instead of the current 65-80%) may leave users and program managers with the false impression of a dramatic loss of skill. Instead, this study defined new benchmarks (Table 2, right) based on the 70th percentile of historical errors at each location and lead-time for the wet-season forecasts." But we have added the following to further emphasize the point "The 70th percentile was chosen because it was relatively close to the overall performance of the current operational benchmarks (see Figure 5)."

6. Figure 1: Legend including basin shape, rivers network (lakes), gauge stations etc should be shown in the figure.
   Legend added as suggested