Response to comments by Anonymous Referee #1 on "Modeling 25 years of spatio-temporal surface water and inundation dynamics on large river basin scale using time series of earth observation data" by V. Heimhuber et al.

We thank anonymous referee #1 for her/his recommendations and useful comments that helped us to improve the quality of our manuscript. In the following, we provide answers to each of referee #1's comments. It is important to mention that the comments of reviewer #1 refer to a previous version of the manuscript of which the layout differs substantially from the published discussion paper. Therefore, we added the page and line numbers that refer to the published discussion paper to each review comment in square brackets so that readers can easily identify what lines of the discussion paper the comments refer to. Questions raised by the reviewer are in bold face, our answers in regular face.

The manuscript entitled "modeling 25 years of spatio-temporal surface water and inundation dynamics on large river basin scale using time series of earth observation data" modeled SWD for 25 years over MDB. In general, this manuscript is well organized and presented. But some minor issues need to be addressed before accepting for publication.

Page 6, line 15-20, [page 11854, line 6-12], This paper would be better to be self-independent and understandable. The authors should be cautious to cite a paper that is not published yet. Please provide more information about the SWE that were calculated in Tulbure and Broich et al. in term of its spatial variability, uncertainties, etc.

We agree with referee #1 that the readers should have a basic understanding of the SWE time series used in this study and the way it was generated so we added additional key features of the product to the related section of the paper

Updated paragraph (page 6, line 13 to 20 [page 11854, line 4-12]):
"*The dependent variable used in this this study is based on a remotely sensed product of SWD. It is a time series of validated, open surface water and flooding extent derived from the seasonally continuous archive of over 25,000 Landsat TM and ETM+ imagery available for the entire MDB from 1986 to 2011 Tulbure and Broich (in review). The methodology for the development of this time series through machine learning based classification of surface water on the imagery is described in (Tulbure & Broich, 2013; Tulbure and Broich, in review). The overall classification accuracy was 99%, with 87% (+/- 3% standard error) and 96% (+/- 2%) producer's and user's accuracy of surface water, respectively. The SWD showed high inter and intra-annual variability across the MDB, with SWD highly vulnerable to hydroclimatic variability (Tulbure and Broich, in review).The SWE time series used Landsat images with ≤ 50% cloud cover (Tulbure and Broich, in review), resulting in times between subsequent observations of SWE from 16 days (Landsat temporal resolution) to a multiple of 16 days.*"

The paper presented here is focused on a new approach for modeling inundation dynamics based on remotely sensed surface water maps and is to be distinguished from deriving of surface water and flooding dynamics such as Tulbure and Broich et al. (in review). The generation of the SWE time series is a large and complex research project in itself and was beyond the scope of this current work, which was to model SWD empirically.

Page 8, line 28, [page 11857, line 2], what's your criteria for using the number of 40% here? Any references to support this number?

The 40% cloud cover threshold per 10x10 km cell was chosen as a tradeoff between maximizing the number of observations used without including cells that were dominated by clouds. This is described in the text (Page 8, line 26) [page 11856, line 26]: The threshold is selected "to preserve a maximum number of valid surface water observations while maintaining acceptable levels of noise and uncertainty in the SWE time series". Without applying a threshold, we would potentially include classified Landsat observations into the analysis, for which a certain modeling grid cell was entirely covered by clouds, despite the fact that a threshold of 50% is already applied in the generation of the SWE time series by Tulbure and Broich et al. (in review). Consequently, the application of another threshold on the level of individual grid cells is inevitable for our application. Lowering the cloud cover threshold leads to an increase in the number of observations used for fitting the models but at the same time, it also leads to higher levels of uncertainty regarding the estimated SWE of each observation. Based on a variety of iterative experiments, we found that 40% is the most suitable value for balancing the loss of data with the uncertainty and noise in the remaining data in the majority of cases. Due to the very limited number of studies that have developed empirical inundation models based on Landsat time series data, there are no references to support the selected threshold value of 40% for the maximum amount of cloud cover that we allow for each Landsat observation per 10 by 10 km grid cell.

Page 9, line 18, [page 11857, line 24], beta 4 should multiply ET instead of P in eq (1)?

Yes. In the published version of the discussion paper, this was already partly corrected, so that beta 3 is multiplying ET and beta 4 is multiplying P. Although the repetition of P was corrected here, P should come before ET in the equation because it is the first variable that is added to the base model in the variable selection process. We updated the equation accordingly and the final equation is:

$$SWE_t = \beta_0 + \beta_1 Lag(Q) + \beta_2 SWE_{(t-1)} + \beta_3 P + \beta_4 ET + \beta_5 SM + e$$

Page 11, line 19, [page 11860, line 11], what is the normal flow concentration time from the upper stream to the lower stream, is 5 days time step small enough to test the correlations? Why not 3 days or 1 day? The authors should adjust the selection of positive and negative of 5 days here.

We agree with referee #1 that in theory, it should be possible to quantify flow lag times based on a finer time step. However, we found that a quantification of flow lag times for intervals of less than 5 days is limited by the following reasons. Firstly, Landsat captures surface water extents ideally every 16 days in most areas and thus only provides limited information about the propagation of floods on a day to day basis. Secondly, the flow travel time between two points along a water course is not entirely static because it is partly a function of discharge. In the paper, we express this issue on page 11 line 20 [page 11860, line 13]: "We selected 5-day intervals to account for the fact that there is no exact lag time for each cell because flow travel times are a function of discharge and overbank flow, so that elevated discharges during times of flooding are likely to result in different flow travel times compared to low flows (Overton et al., 2006)." Consequently, we think that 5 day intervals are an adequate choice for the quantification of flow lag times for 10 by 10 km grid cells across the study area considering the characteristics of the remote sensing data and the above mentioned variable nature of flow travel times.

Page 14, line 10, [page 11863, line 20], should be “directions”

Changed as suggested.

Page 16, line 8, [page 11865, line 27], should be “(Fig. 7)” Figure 1, the unit of the elevation should be in km instead of m.a.s.l

This was already (Fig. 7) in the published version of the discussion paper. In Figure 1, we changed the unit of elevation from m.a.s.l. to km as suggested.

Table 3 and 4 , the unit of discharge (ml/day) seems odd to me, it would be better to convert it into mm/year in table4 then the readers can have a clear sense of the water balance in this area

We agree that the mm/year unit would give a better idea of the water balance but the key issue here is that most of the EH-zones (including the three sub-regions used in this paper), do not cover the runoff generating catchment area. Table 4 is intended to give an overview of the regional climate characteristics for the example zones rather than the water balance. For the water balance, we would have to consider the entire catchment area of the main water course of each zone, which in the case of the Murray zone, covers almost the entire study area. Since we are essentially modeling floods on local floodplain units, we are rather interested in the effect of local rainfall, evapotranspiration and soil moisture on inundation dynamics than in catchment or water balance processes. In other words we wanted to see, whether heavy rainfall events can influence the inundation dynamics on floodplains as observed by Landsat. Based on this consideration, we believe that it is useful to provide spatial averages (mm/year) of local rainfall, evapotranspiration and soil moisture on the level of the example zones. For discharge however, we find that a conversion into mm/year based on the spatial extent of the example zones would be of limited value, since the example zones do not comprise the runoff generating catchment areas (i.e. runoff in the river of a given EH-zone is mainly generated in distant upstream catchment areas).

Since the authors have data available for a longer period, I am wondering how is the model performance for the 2010-2011 La Nina Floods?

Due to the relative complexity of the methodology in this paper, we decided to focus on model development, results and the applicability of the models to different river and floodplain systems rather than model validation or the assessment of the model performance for selected flood events. Although it would be very interesting to assess the models’ ability to predict specific flood events, such an assessment is beyond the scope of this paper. Nevertheless, we give an overview of the performance of the models for each of the three example zones in Table 6, based on the CV-RMSE and r^2 of the fitted models. We chose the CV-RMSE because in 5-fold cross-validation, each 5 year period of the 25 years of data is used for model validation including the 2006 to 2011 period which includes the La Nina Floods. Even though the resulting CV-RMSE does not provide information about the model performance in the specific case of the La Nina Floods, it is still sensitive to the models ability to predict the 2006 to 2011 period based on a model fitted to the remaining data.

Based on the experience that we gained during the analyses for this paper, we can state that the performance of a model for predicting large floods depends on a complex interplay of a variety of factors that are unique for each modeling grid cell. These factors include the hydro-geomorphology of the river and floodplains, the distance to the nearest available river gauge and the level of flow retardation resulting from water management infrastructure such as diversion schemes. As a result of these factors, model performance is highly variable across different grid cells (e.g. Fig. 4c) so that the prediction of specific flood events such as the La Nina Floods is likely going to work well for cells with good model performance (high r^2 and low CV-RMSE) but not as good for cells that are difficult to model due to their specific combination of the above mentioned limiting factors.

References:

Tulbure, M.G. & Broich, M., 2013. Spatiotemporal dynamic of surface water bodies using Landsat time-series data from 1999 to 2011. *ISPRS Journal of Photogrammetry and Remote Sensing*, 79, pp.44–52.