Dear reviewer,

On behalf of my co-authors, we appreciate the time you have taken in our paper and the valuable suggestions that are very helpful to enhance our manuscript. You will find below the answer according to the reviewers' comments.

Reviewer's general comments:

The main goal of the manuscript is to evaluate the ability of the forecasting model ECMWF SEAS5 to simulate the climatology of precipitation and temperature over Malaysia and the accuracy of streamflow predictions, at various lead times, forced by the seasonal atmospheric model for different seasons and sub-regions.

Because of the relevance for water managers, as reinforced in the introduction sections it is clear that the manuscript is of general interest for HESS´s readers.

There are, however, several points in the manuscript that I believe that need to be considered by the authors in order to improve the paper interest.

In my view the manuscript needs to be reduced and focus in analyzing what is relevant in a tropical country, which is rainfall rather than discharge. Then, discussion can concentrate on what matters, which is river discharges. The current version is too long, which makes difficult to read.

This is true that we focus on the hydrological forecast; however, meteorological and hydrology are connected. Moreover, much of the agriculture in the region (e.g. rice) heavily depends on irrigation, not only on rainfall and hydropower is growing sector in all MSEA countries. Both temperature and rainfall are the key components of hydrological system. The VIC model is driven by weather conditions to generate runoff and discharge. Therefore, meteorological forecasts are important to determine the relationship relation or consequence in hydrology part. It is necessary to analyze temperature and rainfall forecasts to provide a more complete hydrological forecast and to get some idea of the propagation of skill through the modelling chain. However, we will improve some figures or leave to supplement. For example, the skill between WFDE and APHRODITE that are similar.

Considering the current limitations of climate model for accurately predicting many month in advance, I wander why the statistics chosen are mainly numerical (thus is, correlation) rather than using categorical indexes of performance.

Thank you for the comment. In our study, we combined both numerical (correlation coefficient and RPSS) and categorical (ROCSS). We first used correlation coefficient and RPSS for general skill assessment and subsequently, we used ROCSS for tercile anomalous year assessment.

Reviewer's specific comments:

Lines 30-35. While I understand the advantage for water managers of using probabilistic forecast at intra-seasonal scales, this paragraph regarding climate change might be out of context because the adverse impacts of global climate change increase along decadal time-scales, while the focus of the study are the forecast for lead times of ~ 30 days.

Thank you for the comment. We put our research in the context of climate change because we would like to state the importance of seasonal forecast system. The increasing climate variability and weather extremes, as a result of changing climate, will impact the hydrological system and other sectors. Therefore, there is an increasing need for seasonal forecast as an early warning system in support of management. We will reformulate this paragraph of our manuscript accordingly, possibly by adding something like "Hence, probabilistic forecasts are necessary as the early warning system to establish a strategy taking uncertainties about future hydrological conditions into account."

Data description item: the manuscript has so many acronyms that makes it hard to remember while reading. My suggestion is to include a table in the data description section indicating time and space resolution of each data set.

This is a good suggestion. We will include a similar table with data descriptions with acronyms (you will find it below) in the supplement.

Table S2 Data description. Please note, that all data sets have the same native resolution (0.5°)

Data	Version	Acronym	Time period
WATCH Forcing Data	ERA5	WFDE5	1983 - 2014
ECMWF ensemble forecast	System 5	SEAS5	1985 - 2014
Temperature APHRODITE	1808		1985 - 2014
Precipitation APHRODITE	1101		1985 - 1997
Precipitation APHRODITE	1901		1998 - 2014

90-95. I am not very familiar with the WFDE5 dataset which was used as a reference data. In my experience, what is relevant in tropical areas for hydrological forecast is to remove the errors in rainfall since precipitation has the greatest impact on discharges. How the WFDE5 dataset compares with rainfall estimations derived, for instance form satellite products such as IMERG, etc. I´ve seen validations were performed against the APHRODITE database, which is based in station data, but apparently not for the whole hindcast period.

WFDE5 is ERA5 reanalysis data, bias-adjusted against CRU observed data on temperature and radiation, and CRU or GPCC observed data on rainfall and no of wet days. It is often used for impact studies, including hydrological and agricultural analyses. The WFDE5 itself has been evaluated against meteorological observation and represents good realistic global hydrological

system. It provides a number of output parameters that are required to run the VIC model in this study. Moreover, the range of available data is long and covers the study period of this study. APHRODITE is the dataset specifically for Asian area, so we used it to validate the WFDE5. However, there are still limitations of using APHRODITE, such as the lack of completeness needed for this study period and not all parameters provided to run the hydrological model are available.

130-135. If I understood correctly, the experiments were performed only for ENSO years. How many events were considered? It is not clear for me whether a single experiment (thus is, a single forecast) for each season was carried out; or several runs for each season were for different initial conditions throughout each season was used to calculate the statistics of performance.

First, we want to stress that all simulations were continuous for **all** the available hindcast years, they were **not** confined to ENSO years only. They were used for monthly skill assessment of figures 3-12.

In additions we analysed skill at seasonally aggregated time scales (MAM, JJA and SON), again continuous for **all** hindcast years, in figures 13-15, but highlighting years that can be classified as positive or negative ENSO phases. This because ENSO is an important driver of climate variability in the MSEA region. We analyzed the ENSO using the R packages to evaluate probabilistic tercile and ROCSS as mentioned in the manuscript. The evaluations were run separately for each season throughout 30 years. As the results shown in figure 13-15, each season will be run for the entire 30 years and show the probability tercile of each year. The ENSO events (black stars or black squares) are plotted on the years that the ENSO occurred during that season.

We will try to better explain in our methods section

145, item 4.2.1 Near surface temperature. The authors need to justify all the statistics involving temperature. While I do understand the hydrological implications of the accuracy of temperature forecasts in a temperate country because it has to do with melting, in my view it is of low interest in a tropical country. Are there any hydrological implications, I mean on discharge values whether the predicted temperature was 28 ^{o}C while the observations was 27^{o} C?

We agree that the temperature variability in the tropic region is low. However, The VIC model was run for the entire hydrological basins in this region (as mentioned in lines 115-116), where the sources of all major rivers are from the Tibetan Plateau. Consequently, the temperature affects the melting and subsequently the hydrological stream flow.

We will add a sub-basins map (the analysis domain for both climate and hydrology) , see example figure below, that is larger than the MSEA area.

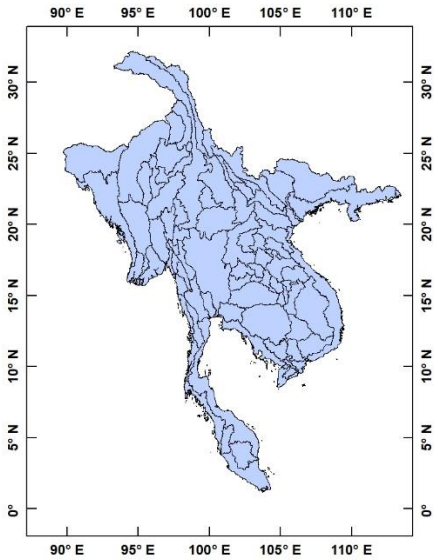
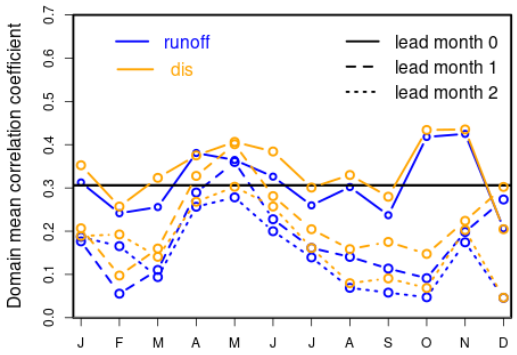


Figure: Subbasins domain map.

160 The fact that the observation driven data set, APHRODITE, “shows 160 a higher skill magnitude compared with the evaluation against WFDE5, especially during the rainy season” might be due to biases in the modeling driven WFDE5 (?).

You make a good point. Generally, the rainfall range (in terms of intensity and frequency) in the tropic region is large. This causes prediction difficulty. The fact that the evaluation with APHRODITE shows a higher skill compared to the evaluation with WFDE5 might be the result that WFDE5 is a model based (though as a reanalysis heavily constrained by data) dataset, and its limited resolution may cause poorer representation of especially rainfall extremes. The bias-correction should reduce some of these errors, but apparently not all. This could be the reason for the slightly larger error in the WFDE5 model. Even though the evaluation with APHRODITE shows a higher skill than with WFDE5, the difference is small.

Figure 7 it appears to me that runoff prediction is better than discharge but the statement of line 217 concluded the opposite.



According to the figure you referred to, the discharge is orange line and runoff is blue line, both representing the correlation coefficient. The discharge and runoff show a similar trend, but it can be seen that the correlation coefficient of discharge shows a little higher score compared to runoff: the orange lines are always above the respective blue lines. Therefore, we conclude that the discharge prediction is higher than runoff.

The problem with this analysis against discharge observation is related to the fact that it includes the uncertainties of the VIC model. Besides all the errors due to the forecasts of the mode, the WFDE5 initialization, etc, how much of the variability can be attributed to the hydrological model itself?

This is a good point and an omission in the paper. It is difficult to define the variability of the VIC model itself because the VIC is data-driven based model. Input data is the main component that give result in variability. We evaluated the correlation coefficient of discharge between real observations (from gauging stations) and WFDE5-driven simulations. You will find the figure below and we will add the figure to the supplement, or combine it with figure 12. The result presents a good correlation coefficient for large parts of the year, so we expect the role of VIC model variability is small in the skill assessments of SEAS5.

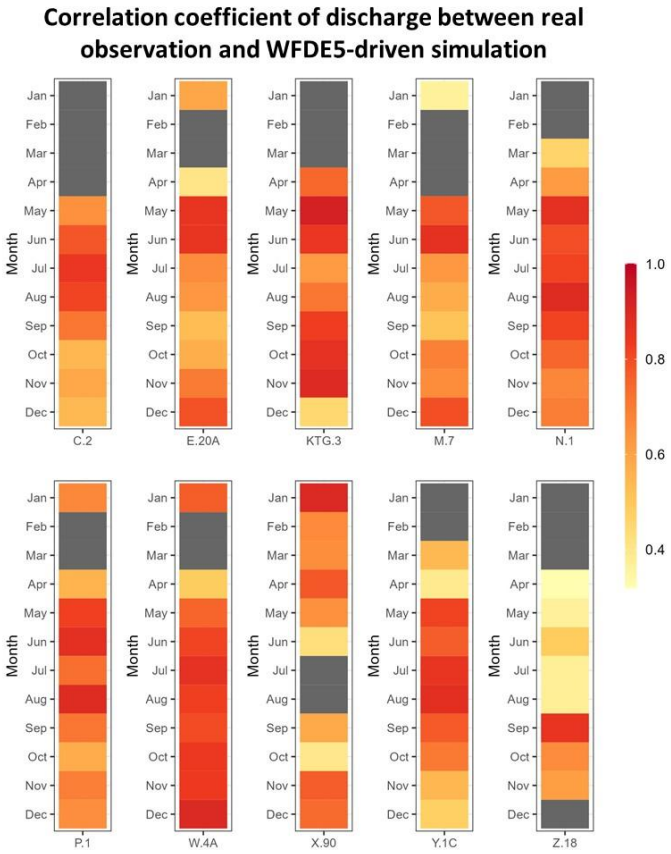


Figure S: Monthly correlation coefficient R ($p < 0.05$) for water discharge generated from VIC model driven by WFDE5 hindcast (reference simulation) against the observation at gauging stations.