Anonymous Referee #2

I have finished my review of the paper "Using NDII pattern for a semi-distributed rainfall-runoff model in tropical nested catchments", by Sriwongsitanon et al., submitted to HESS. This paper outlines a comparison study of four models of the same set of nested catchments in Thailand – a lumped model (FLEXL) applied to individual gauges, the semi-distributed version of the same model (FLEX-SD), FLEX-SD modified by using the NDII remote sensing metric to inform the distribution of soil stores (FLEX-SD-NDII), and the independent semi-distributed URBS model. An attempt is made to demonstrate (1) the improved accuracy/realism of using NDII to inform the spatial distribution of soil stores while only calibrating a reference storage quantity and (2) the superiority of FLEX variants over the URBS model. This short paper is generally well-written but I have found it to include critical experimental design issues, flawed interpretation of results, and procedural omissions. I therefore believe that it is not currently acceptable for publication in HESS, and recommend rejection. I outline the reasons for this below.

Major Comments:

1) The authors only report calibration statistics and performance for their models. There is no effort to validate the models using even standard split-sample validation. Calibration without validation is no longer generally deemed acceptable practice for evaluating hydrological model performance or individual model choices and renders most of the paper results not particularly convincing. In particular, the arguments that the NDII informed model is "more realistic" due to improved performance in calibration alone are unjustified.

Answer:
We are very grateful for the detailed reading by the referee#2 and his/her valuable criticism and suggestions. However, we fear that there is some misunderstanding on the purpose of our study, which we apparently did not make sufficiently clear. We respectfully disagree with the referee on the need to do split-sample validation for our model results to be realistic. First of all, this is not a calibration study, but a study to show that a model which is calibrated on a single outfall, can be realistically applied to compute the runoff behavior in nested catchments based on independently available topographic (catchment size and river length) and RS (NDII) information. The validation is done by comparing the outcome with runoff observations at internal stations in a number of nested catchments and with independently derived information on Soil Wetness (SWI). This is a far better approach to validate model realism than by using split-record tests, where the validation seldom gives very different results as the calibration. The calibrated FLEXL models are merely used as bench marks for comparison to the results obtained by the single catchment-wide FLEXL-SD model. This paper is not about showing that the FLEXL models can be successfully calibrated and applied, it is used to show that a lumped model for a large catchment can be used to set-up a semi-distributed model that can compute runoff in the entire network of nested sub-catchments, making use of additional external information.

The realism of this approach is not derived from the performance statistics, but 1) from how well this approach is able to predict runoff at internal stations compared to both observations and calibrated lumped models of sub-catchments, and 2) from how well this approach is able to

simulate root zone soil moisture, a crucial internal state variable, compared to independently derived SWI values.

We do not claim that "the NDII informed model is more realistic due to improved performance in calibration alone". On the contrary. We make this claim because we validate the performance against independent information (not used in calibration) of internal runoff stations and of RS derived SWI (also not used in calibration). We also show that the FLEX-SD-NDII compares better with SWI values than the individually calibrated FLEXL models of the catchments.

2) At multiple points in the paper, the authors report that the model "gained realism" (e.g., line 15 & 17 of pg 10)– I look at figures 4 and A2 and see only an improvement in baseflow simulation in basins P.20 and P.21 (the only headwater basins unaffected by the Mae Ngad dam overwriting of flows); this is consistent with the quantitative KGE_L metrics which are much more objective assessment of model skill. However, the KGE_L metrics denote a degradation of baseflow in 3 other non-headwater basins – is this still therefore a gain in model realism? Here, the authors can discuss hydrograph fit (in calibration conditions only) but there is no evidence that the mode "gained realism", and I'm rather sure that this is not something that could ever be determined via observation of a hydrograph alone. This is the primary contribution of section 5.1 and it is not defensible from the experimental data. Interpretation of these results seem cherry-picked. Alternate interpretations of the data in table 4 and figures 3,4,5 likewise denote quite inconsistent performance of the FLEX-SD-NDII except for low flow in the 2 headwater basins. Had the authors calibrated the FLEX-SD model using weighted calibration all gauges of interest rather than just P.1 I expect they could get comparable performance without NDII; this likely should have been another test model configuration.

Answer: Again, the referee is mistaken in that this is not a calibration study. Of course we could have used all the runoff stations in the calibration of the entire catchment. But this was not the purpose. It was not our purpose to improve the calibration of the Upper Ping basin, but to show that on the basis of a calibrated large basin (at P.1) a semi-distributed model can be set up, using only topographical information and a readily available RS-derived indicator, to simulate the runoff at any point within that basin with good results. "Good results" here means that they compare favorably (as good or better) to calibrated sub-catchment models (which in this case were available, but not in a PUB case) and to an independently derived distributed moisture state indicator. This paper aims to make a contribution to prediction in ungauged basins (PUB) and not to the calibration of models, even though the FLEX-SD-NDII model is better capable of representing moisture states than the FLEXL for the entire basin.

Purely for reasons of comparison, we also calibrated FLEX-SD and FLEX-SD-NDII at each gauging station. Table 1 presents the statistical indicators at each station using 4 models calibrated at these stations comparing them to the predictions based on calibration at P.1 only. Obviously the models when calibrated at the sub-catchment stations generally perform better, but in those cases additional runoff information at these particular stations was used. For the models calibrated on P.1 only, no information on internal fluxes were used, yet the performance is not much worse. Hence this approach shows that FLEX-SD-NDII is a valuable approach to predict internal fluxes when no discharge observations at nested catchments are available.

Table 1. Statistical indicators at each station simulated using 4 models by the calibration at each station compared to the calibration at P.1

Station	Model - Case		NSE	KGE_E	KGE_L	KGE_F
P.1	(1) URBS	Calibrated at P.1	0.83	0.91	0.81	0.98
	(2) FLEXL	Calibrated at P.1	0.85	0.92	0.73	0.98
	(3) FLEX-SD	Calibrated at P.1	**0.88**	**0.94**	**0.81**	**0.99**
		Calibrated at P.1	**0.88**	**0.94**	**0.81**	**0.99**
	(4) FLEX-SD - NDII	Calibrated at P.1	**0.88**	**0.94**	0.72	**0.99**
		Calibrated at P.1	**0.88**	**0.94**	0.72	**0.99**
P.4A	(1) URBS	Calibrated at P.1	0.72	0.85	0.25	0.94
	(2) FLEXL	Calibrated at P.4A	**0.77**	**0.88**	0.70	0.97
	(3) FLEX-SD	Calibrated at P.1	**0.77**	0.87	0.73	0.94
		Calibrated at P.4A	0.76	**0.88**	**0.76**	**0.99**
	(4) FLEX-SD - NDII	Calibrated at P.1	0.76	0.86	0.47	0.95
		Calibrated at P.4A	0.76	**0.88**	0.75	**0.99**
P.20	(1) URBS	Calibrated at P.1	0.60	0.57	-0.38	0.62
	(2) FLEXL	Calibrated at P.20	**0.67**	0.83	0.71	0.98
	(3) FLEX-SD	Calibrated at P.1	0.63	0.53	0.44	0.56
		Calibrated at P.20	**0.67**	**0.84**	**0.77**	**0.99**
	(4) FLEX-SD - NDII	Calibrated at P.1	0.62	0.62	0.68	0.67
		Calibrated at P.20	**0.67**	0.83	0.64	**0.99**
P.21	(1) URBS	Calibrated at P.1	0.61	0.74	-1.36	0.79
	(2) FLEXL	Calibrated at P.21	0.73	0.86	0.85	**0.99**
	(3) FLEX-SD	Calibrated at P.1	0.56	0.72	0.39	0.76
		Calibrated at P.21	0.76	**0.88**	0.76	**0.99**
	(4) FLEX-SD - NDII	Calibrated at P.1	0.61	0.77	0.64	0.86
		Calibrated at P.21	**0.77**	**0.88**	**0.88**	0.97
P.67	(1) URBS	Calibrated at P.1	0.78	0.85	0.76	0.90
	(2) FLEXL	Calibrated at P.67	0.80	0.90	**0.80**	**0.99**
	(3) FLEX-SD	Calibrated at P.1	**0.83**	0.86	0.74	0.89
		Calibrated at P.67	0.82	**0.91**	0.75	**0.99**
	(4) FLEX-SD - NDII	Calibrated at P.1	0.82	0.84	0.64	0.87
		Calibrated at P.67	0.82	**0.91**	0.77	0.98
P.75	(1) URBS	Calibrated at P.1	0.71	0.82	0.80	0.87
	(2) FLEXL	Calibrated at P.75	0.74	0.86	0.71	0.96
	(3) FLEX-SD	Calibrated at P.1	0.78	0.85	0.82	0.89
		Calibrated at P.75	**0.80**	**0.90**	**0.83**	**0.99**
	(4) FLEX-SD - NDII	Calibrated at P.1	0.76	0.84	0.79	0.88
		Calibrated at P.75	0.79	**0.90**	0.82	0.98
Average	(1) URBS	Calibrated at P.1	0.71	0.79	0.15	0.85
	(2) FLEXL	Calibrated at each station	0.76	0.88	0.75	0.98

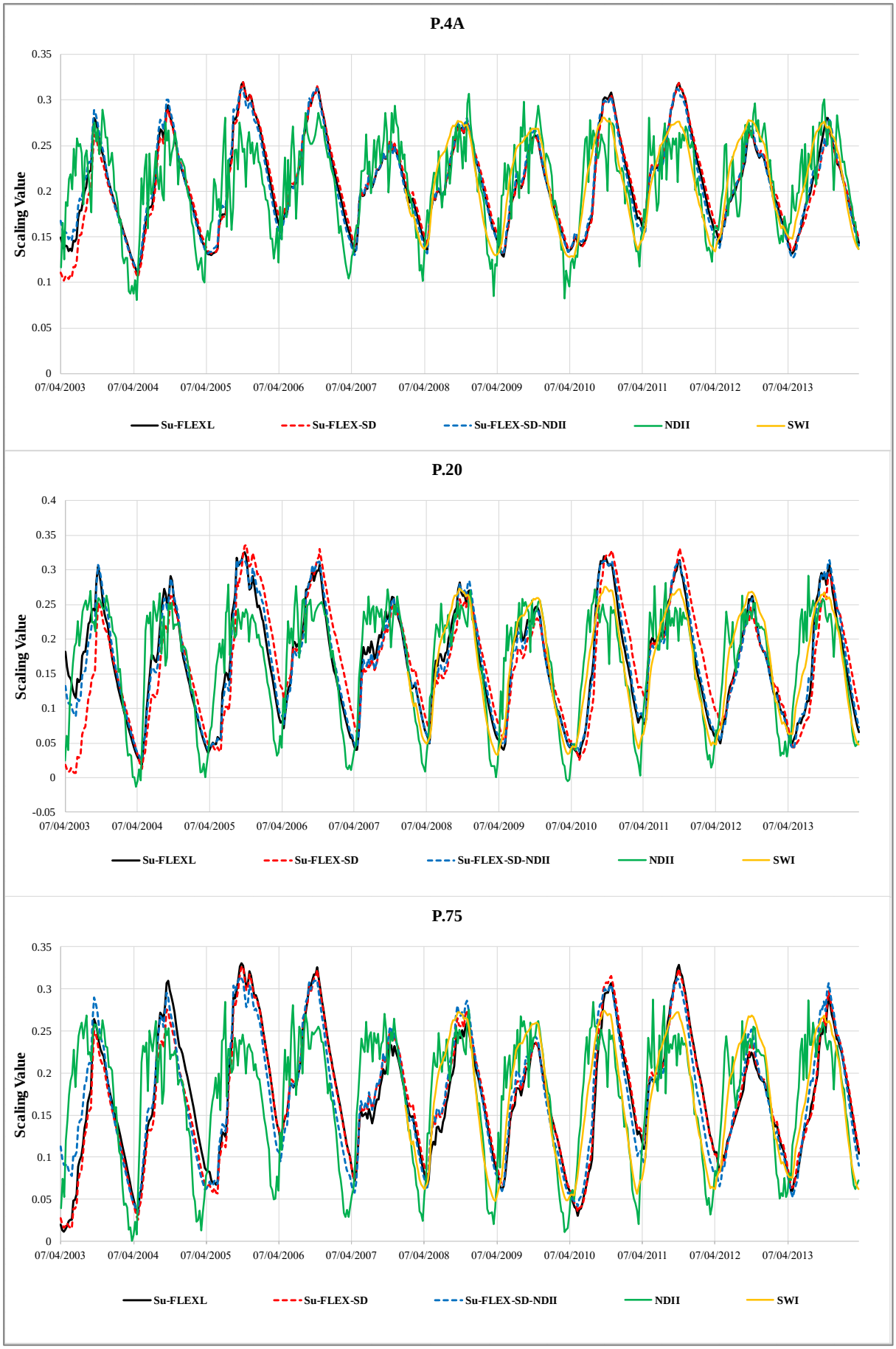
(3) FLEX-SD	Calibrated at P.1	0.74	0.79	0.66	0.84
	Calibrated at each station	**0.78**	**0.89**	**0.78**	**0.99**
(4) FLEX-SD - NDII	Calibrated at P.1	0.74	0.81	0.66	0.87
	Calibrated at each station	**0.78**	**0.89**	0.76	0.98

3) The comparison of the correlations of SWI and soil storage characteristics in figure 7 (the other primary piece of evidence supporting the improvement incurred by distributing soil storage information) is likewise unconvincing. All models exhibit the same trend, and the moderate improvements in R^2 metrics for a cloud of points about a power law curve (especially when the data relation does not look to have the shape of a power law) is not sufficient to demonstrate model improvement. Again, this is particularly true in light of the fact that all evaluations are done during the calibration period.

Answer:

We are not trying to improve the accuracy of model performance by developing the semi-distributed models, however we are focusing on simulating runoff estimates at required locations upstream of a calibrating station with an accuracy similar to the results provided by the lumped model which requires model calibration at all stations.

To demonstrate the relationships of Su-NDII and Su-SWI, instead of using the R^2 values from the scatter plots as shown in the manuscript, we compared the correlation between the time-series of Su from 3 models compared to the time series of SWI and NDII for 6 stations (see **Figure 1** below). Since the scale of these 3 variables are different, a scaling algorithm was applied. The results of R^2 in **Table 2** and **Table 3** show that the values of SWI correlate well with Su not only in the dry season, like for the case of NDII, but also in the wet season. Except for P.20, the Su simulated with FLEX-SD-NDII appears to have a higher correlation with SWI than the ones provided by FLEXL and FLEX-SD.


P.4A
Scaling Value
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
07/04/2003
07/04/2004
07/04/2005
07/04/2006
07/04/2007
07/04/2008
07/04/2009
07/04/2010
07/04/2011
07/04/2012
07/04/2013
Su-FLEXL
Su-FLEX-SD
Su-FLEX-SD-NDII
NDII
SWI
P.20
Scaling Value
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
07/04/2003
07/04/2004
07/04/2005
07/04/2006
07/04/2007
07/04/2008
07/04/2009
07/04/2010
07/04/2011
07/04/2012
07/04/2013
Su-FLEXL
Su-FLEX-SD
Su-FLEX-SD-NDII
NDII
SWI
P.75
Scaling Value
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
07/04/2003
07/04/2004
07/04/2005
07/04/2006
07/04/2007
07/04/2008
07/04/2009
07/04/2010
07/04/2011
07/04/2012
07/04/2013
Su-FLEXL
Su-FLEX-SD
Su-FLEX-SD-NDII
NDII
SWI

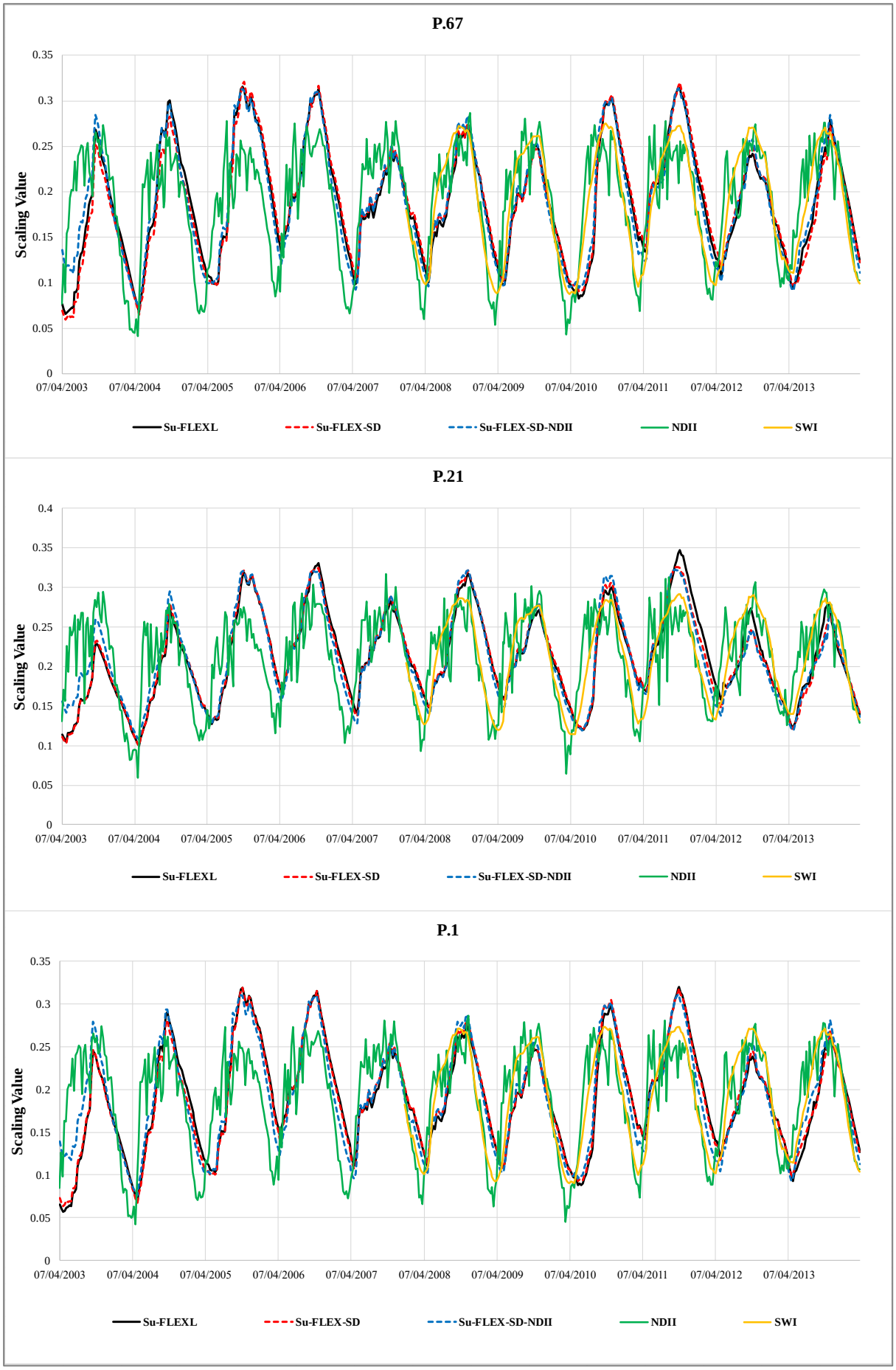

P.67
Scaling Value
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
07/04/2003
07/04/2004
07/04/2005
07/04/2006
07/04/2007
07/04/2008
07/04/2009
07/04/2010
07/04/2011
07/04/2012
07/04/2013
Su-FLEXL
Su-FLEX-SD
Su-FLEX-SD-NDII
NDII
SWI
P.21
Scaling Value
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
07/04/2003
07/04/2004
07/04/2005
07/04/2006
07/04/2007
07/04/2008
07/04/2009
07/04/2010
07/04/2011
07/04/2012
07/04/2013
Su-FLEXL
Su-FLEX-SD
Su-FLEX-SD-NDII
NDII
SWI
P.1
Scaling Value
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
07/04/2003
07/04/2004
07/04/2005
07/04/2006
07/04/2007
07/04/2008
07/04/2009
07/04/2010
07/04/2011
07/04/2012
07/04/2013
Su-FLEXL
Su-FLEX-SD
Su-FLEX-SD-NDII
NDII
SWI

Figure 1. Time series of Su from 3 models compared to NDII and SWI of 6 stations

Soil moisture Index	Model	P.4A	P.20	P.75	P.67	P.21	P.1
NDII	FLEXL	0.43	0.48	0.39	0.40	0.39	0.38
	FLEX-SD	0.39	0.39	0.40	0.39	0.37	0.39
	FLEX-SD-NDII	0.44	0.47	0.46	0.45	0.41	0.45
	Average	**0.42**	**0.45**	**0.42**	**0.41**	**0.39**	**0.41**
SWI	FLEXL	0.80	0.89	0.77	0.79	0.72	0.75
	FLEX-SD	0.77	0.77	0.78	0.77	0.70	0.77
	FLEX-SD-NDII	0.83	0.87	0.87	0.84	0.73	0.83
	Average	**0.80**	**0.84**	**0.81**	**0.80**	**0.72**	**0.79**

Table 2. R^2 values for SU-NDII and SU-SWI relationships during the wet season of 6 stations

Soil moisture Index	Model	P.4A	P.20	P.75	P.67	P.21	P.1
NDII	FLEXL	0.72	0.81	0.69	0.75	0.53	0.68
	FLEX-SD	0.69	0.67	0.70	0.70	0.56	0.69
	FLEX-SD-NDII	0.77	0.78	0.78	0.77	0.62	0.76
	Average	**0.73**	**0.75**	**0.72**	**0.74**	**0.57**	**0.71**
SWI	FLEXL	0.82	0.91	0.78	0.83	0.57	0.78
	FLEX-SD	0.79	0.74	0.78	0.79	0.58	0.78
	FLEX-SD-NDII	0.87	0.86	0.85	0.86	0.63	0.85
	Average	**0.83**	**0.84**	**0.80**	**0.83**	**0.59**	**0.80**

Table 3. R^2 values for SU-NDII and SU-SWI relationships during the dry season of 6 stations

4) The authors attribute (at line 25 of pg 9) model discrepancies to observed due to flow regulation at the Mae Ngad Dam upstream of P67. However, they state at line 3 of page 4 that reservoir outflow data was explicitly used as input to model. This flow regulation should therefore be perfectly handled within the model (and in fact would inflate model fit statistics, since the reservoir flows would comprise a rather large portion of the hydrograph, especially under low flow conditions).

Answer:

We will revise the manuscript since reservoir outflow data was used as input to the models. However, regulated flows for irrigation are usually present within this catchment. It could affect the overall performance of each model since the regulated flows are not recorded to be included in the model simulation.

5) The reporting of the calibration process is inadequate. What was the calibration period? What was the objective function? How did the authors separate low flow statistics (KGE_L) and high flow statistics (KGE_E)? Was a run-up period used? Why not? Is this hourly NSE or daily NSE? Was the model run at an hourly time step as implied by the use of hourly time lags? The MOSCEM optimization algorithm is uncited. There were a significant number of critical details missing that ensure that these experiments are not replicable.

Answer:

Rainfall and runoff data are available between 2003 and 2013, we calibrated the models during this period. The objective functions are the Kling-Gupta Efficiencies for high flows, low flows, and the flow duration (KGE_E, KGE_L and KGE_F), respectively. KGE_E is analyzed using the following equations. KGE_L can be calculated using the logarithm of flows to emphasize low flows.. The model calculates at daily time steps, but this is disaggregated to hourly to take into account the time lags. The output is again aggregated to daily time steps.
We will improve all missing points in the revised manuscript.

$$KGE = 1 - ED \quad (1)$$

$$ED = \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (2)$$

$$\alpha = S_Y / S_X \quad (3)$$

$$\beta = \bar{Y} / \bar{X} \quad (4)$$

$\bar{X}$	=	the average observed discharge
$\bar{Y}$	=	the average simulated discharge
S_X	=	the standard deviation of observed discharge
S_Y	=	the standard deviation of simulated discharge
r	=	the linear correlation between observations and simulations

6) The value of including the URBS model in this comparison is unclear. All in all, I believe that the paper's approach for distributing soil storage capacities using information gleaned from NDII may have merit, but the experimental design did not clearly demonstrate that this approach actually works for the reasons above. While it was demonstrated via the data that moving from a lumped to a distributed approach somewhat improved model performance (which is not entirely surprising, as additional routing information is included with additional free parameters), this is not new.

Answer:

The main objective of developing the semi-distributed model in this study is not for improving the accuracy of model performance but for providing flow estimates at any sub-catchments upstream of the calibrating station. This study also proves that FLEX-SD-NDII can provide simulated flows at any required locations with a similar degree of accuracy compared to simulated flows provided by FLEXL which involves model calibration at each required stations.

Minor Comments:

I have not included the very many minor adjustments I would suggest if I suggested revision, but have touched upon the larger minor issues.

1) In this short 11 page - paper, 2.5 pages (plus 2 full page tables and a figure) were dedicated to defining the models already documented elsewhere in Sriwongsitanon et al. 2016 and Carroll 2004. A simple reference would do for most of these details
(especially for snow simulation in Thailand!)

Answer: Thank you for your point. On the other hand, we have also had the comment that a paper should be readable by itself, without having to look up the description in another paper. This is why we briefly summarized the model structure and equations.

2) In presentation of calibrated parameters of table 5, mixing actual calibrated parameters with those calculated from area scaling (Tlag) and the NDII relation (Sumax). The catchment wide Sumax (a calibration parameter) not reported. For some reason the Sumax_i from all of the basins with FLEX-SD-NDII are all less than FLEX-SD, which struck me as odd.

Answer: FLEX-SD and FLEX-SD-NDII were calibrated only at P.1 station. Therefore, the parameter values are only presented at P.1 except TlagF and TlagS for both models and Sumax for FLEX-SD-NDII. Indeed the values of Sumax_i are smaller than the Sumax value of FLEX-SD at P.1. This shows how tricky it is to attribute physical meaning to lumped parameter values in a situation where parameters can compensate for each other (particularly Sumax, Ce, Beta and D). Again, this paper is not about calibration, but on how additional independent information can be used to disaggregate overall catchment performance to nested sub-catchments. The details on performance and on parameters are provided to analyse how this is achieved, and not to prove that this method is better than individual calibration of nested catchments.

3) observations of improved performance of NDII model @ pg 10 ln 1 not consistent with KGE_L reporting for same basin

Answer: The realism of a model result does not merely depend on performance indicators. These indicators are fine for a quick screening or filtering of behavioral parameter sets. However, close scrutiny of the performance of hydrographs, which demonstrate the detailed dynamics of the model compared to observations, is far more telling. In this case, we can see clearly in Figure 4 of the manuscript that FLEX-SD-NDII simulations (the green lines) follow the pattern much more closely, especially during low flows. One would expect KGE_L to then also provide a higher value. Why this is not the case may be due to the fact that Figure 4 plots on logarithmic scale.

4) in table 4, the “best performance is underlined”, however, this is not the case, as FLEXL often has best performance.

Answer: Again, it is not our intention to show that the FLEX-SD-NDII for each sub-catchment performs better than the individually calibrated FLEXL models. Here it is the intention to show which of the SD models (calibrated only on P.1) shows the best performance. FLEXL and URBS are mentioned merely as a reference.

5) Figure three has little value – a percent bias reporting would be more succinct and equally valuable.

Answer: We will take the mass curves out and replace with percent bias in the revised manuscript.