We thank the referee for reviewing our manuscript and for providing valuable suggestions. Our responses to specific comments are given below. Please refer to the list of symbols in the paper as needed.

GENERAL COMMENTS The authors present a methodology to account for heterogeneity in the calibration of the curve number method (CN) from data. The focus of the study is understanding the variation of CN as a function of precipitation by analyzing the variability of initial infiltration over the catchment. I particularly liked the analysis of the inconsistency of the theoretical definition of initial abstraction (Ia) and its value at the watershed scale. Based on their analysis, the authors propose a set of models with increasing complexity. They apply these models to synthetic basins with controlled heterogeneity following the CN behaviour and compare their performance with standard indices. By introducing additional parameters in the CN method they obtain a good fit of the precipitation-runoff relationship resulting from the application of the CN method to heterogeneous basins.

The topic is relevant for the audience of Hydrology and Earth System Science, because the CN is the most widely used method to account for infiltration losses in professional applied hydrology. The objectives of the study are clearly identified, the methodology for the analysis is sound and the conclusions are relevant and correctly supported by the results and discussion. The proposed models perform well when reproducing the behaviour of heterogeneous basins and there are reasonable expectations that the method can be applied to natural basins. Therefore, I believe the paper deserves publication in Hydrology and Earth System Science.

SPECIFIC COMMENTS I also think that there are several aspects of the paper that deserve a deeper discussion, such as the following:

Referee's Comment

a) On page 2, lines 8-9, the authors state that, in addition to varying spatially due to watershed heterogeneity, CN also varies temporally due to changes in soil moisture or vegetation cover. However, in their synthetic experiments they did not account for temporal variation of CN or Ia. In my opinion, this is a significant limitation for the practical application of the proposed models, that were tested under steady conditions. The authors cite a forthcoming paper by themselves (Santikari and Murdoch, 2018) where several ways of dealing with temporal variation of CN are proposed. I was not able to locate such paper in HESSD. I think this issue should be briefly discussed in this paper, maybe in a section devoted to the limitations of the methodology presented here.

Authors' Response

We agree that not accounting for temporal variations is a limitation of the models proposed. This is why we extended our work to develop models that account for temporal variations, and we present them in the companion paper (Santikari and Murdoch, 2018), which is under review in Water Resources Management. We provided that manuscript during our original submission of this paper to HESS as "author's response", but it appears that it was not accessible. We would like to give the referees access to the companion manuscript, but we are unaware of how to do so. We apologize for the inconvenience.

The focus of the current paper is the inclusion of spatial variations. We first show that the poorly understood variation of CN with *P* is due to watershed heterogeneity. Then we propose models that treat the CN method's parameters as functions of *P*, to account for spatial variations and improve runoff predictions. Although accounting for temporal variations further improves runoff predictions (Santikari and Murdoch, 2018), the proposed models in the current paper have two advantages: (i) they are simpler, and (ii) their data requirements are lower. Models that account for temporal variations are relatively more complex, and they require continuous rainfall-runoff observations for calibration (Santikari and Murdoch, 2018). So, the models proposed in this paper are applicable when continuous observations are absent, and they may be preferable in some cases because of their simplicity.

As per the referee's suggestion, we added a subsection to state the model limitation as follows:

6.5. Model Limitation

A strength of the models proposed in this paper is that they provide a compact way to account for the spatial variation of CN, I_a, or S (watershed heterogeneity), but a limitation is that they do not account for the temporal variation. During dry periods I_a, and S increase whereas CN decreases. The behavior is opposite during the wet periods. Changes in land cover introduce additional temporal variations. Therefore, the calibrated model parameters in this paper can be considered as temporal averages. The models may underpredict runoff during wet periods and overpredict during dry periods. A procedure to account for temporal variations using antecedent moisture is described in the companion paper (Santikari and Murdoch, 2018).

Referee's Comment

b) The promised paper (Santikari and Murdoch, 2018) will also deal with application to real watersheds, not synthetic data (lines 23-26). The argument given in favour of using synthetic watersheds (page 30, lines 12-16) is sound. However, the strength of the proposed CN-based methods lies on their practical applications. Since the authors have analysed applications to real watersheds, I think a brief discussed of this issue should also be included in this paper.

Authors' Response

As per referee's suggestion, we added a subsection to briefly discuss the results from the application of the proposed models to real watersheds as follows:

6.2. Application to Real Watersheds

The models were also evaluated using rainfall-runoff observations from 9 real watersheds located in different parts of the world (Santikari and Murdoch, 2018). Models' ability to predict the observed runoff was assessed using NSE_Q. Results show that in all the watersheds VIMs performed better than CMs but the difference in performance, ΔNSE_Q, varied across the watersheds. Between VIMλ and CM0.2, $\Delta NSE_Q < 0.05$ in one watershed, $0.05 \leq \Delta NSE_Q < 0.7$ in 6 watersheds, and $\Delta NSE_Q \geq 0.7$ in 2 watersheds. Between VIMλ and CMλ, $\Delta NSE_Q < 0.05$ in 3 watersheds, $0.05 \leq \Delta NSE_Q < 0.1$ in 4 watersheds, and $\Delta NSE_Q \geq 0.1$ in 2 watersheds. Based on their performance, the models can be arranged from the best to the worst as VIMλ > VIMS > CMλ > CM0.2, which is consistent with results from their application to the synthetic watershed.

Note: The current subsection "6.2. Model Suitability" has been moved to 6.4.

Referee's Comment

c) On page 11, lines 17 to 21, the authors report the standard professional practice of accounting for heterogeneity by obtaining the area-weighted average of the CN. The results presented in the paper clearly show that this practice can be improved. I think the authors should discuss this in the final part of the paper. This practice is routinely applied in ungauged basins, where CN is estimated from physiographic characteristics. Are there any better alternatives for computing an average CN in view of the research carried out? Can they propose a model for ungauged basins? I am aware this is not the main objective of the work, but I think the paper would benefit from a discussion of this issue.

Authors' Response

We thank the referee for raising this important issue. The alternative to using an average CN is to use an average *Q*, i.e. calculate runoff from each HRU and take the area-weighted average to get the runoff from the watershed. This procedure accounts for heterogeneity (the spatial variation of CN). If watershed scale Ia and CN are estimated from this area-weighted *Q*, they will vary with *P* as shown in Figure 2. Averaging CN is easier to use but averaging *Q* is more accurate.

Throughout the paper, we referred to this approach of averaging *Q* as distributed parameter CN model. We acknowledged in the introduction (page 2, lines 3 and 4) that this approach can account for heterogeneity when it is known at sufficient detail. In section 2.1.1, we mentioned that averaging *Q* is a better approach than averaging CN (page 9, lines 8 to 11). We also used this approach to generate synthetic runoff (Section 5.1: page 23, lines 5 to 7), which was used in the evaluation of the lumped parameter models. We feel that the information provided in the paper makes a strong case in favor of averaging *Q* over averaging CN.

Referee's Comment

d) The models were tested just for one synthetic watershed (described in table 2). This is a limitation of the methodology. The comparative results of model performance would certainly depend on the degree of heterogeneity of the tested basin. In suggest that a discussion of this issue be included in the paper and acknowledged in the conclusions.

Authors' Response

Although we presented results for one synthetic watershed, we tested the models for several distributions of heterogeneity. The summary of our findings was that CM0.2 was always the worst whereas VIMs were better or equal in performance to CMλ. As the referee correctly pointed out, the difference in performance between VIMs and CMλ depends on the degree of heterogeneity. To illustrate this, we evaluated the models for a different distribution of heterogeneity and presented the results in a new subsection as follows.

6.3. Effect of Degree of Heterogeneity

The degree of heterogeneity, defined as the sharpness of change in CN, I_a, or S between the HRUs, may affect the relative performance of the models. To verify this, the degree of heterogeneity of the synthetic watershed (Table 2) was increased by doubling the S_is for HRUs 3 and 4 while the others were left unchanged, i.e. the modified distribution was S_0 = 0 mm, S_1 = 50 mm, S_2 = 100 mm, S_3 = 300 mm, and S_4 = 400 mm. The models were applied to this modified synthetic watershed, for the cases of λ_i = 0.2 and 0.5, and their performances were assessed using SEE_Q.

Comparing the results (Tables 3 and 4) shows that the performance of VIMs remained nearly the same, whereas the performance of CM0.2 decreased and that of CMλ increased. The relative order of performance remained unchanged, i.e. VIMλ > VIMS > CMλ > CM0.2.

Table 4. Performance of the models for the cases of $\lambda_i = 0.2$ and 0.5, assessed using SEE_Q, when the degree of heterogeneity in the synthetic watershed (Table 2) was increased by doubling the S_is for HRUs 3 and 4.

Model	$\lambda_i = 0.2$	$\lambda_i = 0.5$
CM0.2	1.54	1.30
CMλ	0.19	0.38
VIMS	0.12	0.25
VIMλ	0.06	0.12

The results from real watersheds (Santikari and Murdoch, 2018) also show that the performance of CM0.2 was poor, $NSE_Q < 0.25$, in watersheds with a sharp change in CN. Therefore, CM0.2 appears to be unsuitable when the degree of heterogeneity is large. CMλ performed moderately well on synthetic and real watersheds with a large degree of heterogeneity, possibly by transferring the storage (Section 6.1). So, CMλ is suitable for predicting overall runoff, but less reliable for predicting heterogeneity or runoff from small events. VIMs outperformed CMλ in synthetic (Table 4) as well as real watersheds (Santikari and Murdoch, 2018) with a large degree of heterogeneity, and therefore they are more reliable.

Note: The results in Table 4 were obtained from a synthetic watershed that is different than the original (Table 2). Similar results can be obtained by modifying the HRU distribution in Table 2 and reapplying the models. We chose not to present results from several synthetic watersheds because of the similarity of the results (i.e. same relative model performance) and space limitations.

Referee's Comment

TECHNICAL CORRECTION From the formal standpoint, the paper is very well written, correctly organized and adequately illustrated with tables and figures. Figures 8 and 9 could benefit from the use of colours, if possible. Although I am not a native English speaker, I believe the following expression should be corrected: On page 32, line 12 and 14, in both the cases...... ("the" should be removed?).

Authors' Response

We appreciate the comments on the writing and organization in our paper. We agree that figures 8 and 9 would benefit from color, and we used colors in the original submission. It appears that the file that was reviewed may have been converted to black and white. We also noted that the line and page numbers used by the referee are different from those in the copy we submitted, which is same as the pdf file that can be currently downloaded from the HESS website. We are unsure of how and when the file conversion has occurred. The line and page numbers we used in this response correspond to the pdf file that can be currently downloaded. We apologize for the inconvenience.

We thank the referee for pointing out the grammatical error, "the" has been removed.