We would like to thank Jan Seibert and an anonymous reviewer for their constructive comments, which has lead to further analysis and a general improvement of the manuscript.
Below, our replies to each comment are given in blue.

J. Seibert (Referee)
jan.seibert@geo.uzh.ch
Received and published: 12 April 2012
This is an interesting study on the importance of corrections of precipitation measurements for hydrological modelling. Often this issue is not addressed in any detail but the assumption is, that these measurement errors are implicitly corrected in the calibration of a hydrological model. Therefore, it is then assumed that rough corrections are enough or that corrections are not needed at all. This study demonstrates, that these often used approaches might not be appropriate and that we should spend more efforts in better correcting measurement errors. The recommendation, which follows from this study, namely that national weather services should provide (also) corrected precipitation data, seems important. Below I list a number of comments which I hope will help to further improve this valuable study. Some of the comments would require additional computations. While I am aware that such a request is seldomly appreciated, I feel that these additional test would be highly valuable to make this study even more useful.

Methods :
1) Please describe the correction methods better. On pages 3611/12 it remains, for instance, unclear where the snow fraction parameter alpha comes from,

That is explained further in section 2.2, which is expanded in the revised manuscript.

The alpha values are based on a formulation by Vejen (2005) and estimated based on the temperature. This is clarified in the revised manuscript.

"snow fraction parameter α, which is considered a function of temperature, see details on α in section 2.2.."

" • Discrimination between liquid and solid precipitation is performed based on temperature. Liquid > 2 °C ($\alpha = 0$), solid < 0 °C ($\alpha = 1$), mixed precipitation between 0 and 2 °C ($\alpha = 1-0.5 \cdot T$) following Vejen (2005)."

what the reference height for wind speed is

1.5 m

, and how and over which time the different variables

are aggregated (daily, hourly, only during rainfall ? Weighted for rainfall amounts ? . ..).

For the Allerup model, the precipitation was measured every 12 hours, and wind speed and temperature were averaged over this time interval. For our application, which is based on daily precipitation, all variables are averages/sums over 24 hour.

Please also clarify on which data the empirical factors are based.
It is written that: "The Nordic test facility was established in Jokionen, Finland (1987–1993). Based on data from the Finnish test site, the correction method.."

2) While the simplifications on page 3613 sound reasonable, I would like to see some more motivations/justifications on these different simplifying decisions. I am sure you spend a lot of thinking about these decisions, but as this part reads now it sounds a bit ad hoc.

Of course there is an issue of data availability, since wind speed and temperature is only available either as sparse point measurements or on a daily scale. However, precipitation is also only available as daily sums, meaning that wind speed and temperature during precipitation cannot be obtained. Based on previous publications (Vejen, 2005) and especially (Allerup et al, 2000) who examined the effect of using off-site measurements, there is reason to believe that the spatial interpolation does not cause a major problem when applying the 20 km grid data for wind and temperature. Monthly precipitation intensity values could also be problematic to apply, but since intensity only occurs in the correction of liquid precipitation, which has much smaller correction factors, this is not believed to be a significant water balance issues (see. E.g. figure 3, where summer precipitation is almost identical between the two correction methods).

The justification of simplifications is expanded in the revised manuscript.

Vejen 2012 (personal communication) is currently calculating updated correction factors for the period 1990-2012 based on 12 climate stations and report almost identical total precipitation volumes as reported in this study (figure 1) based on the simplifications described. The new calculations by Vejen are however not published or available yet.

Modeling
3) Much could be discussed about whether the SHE model actually can, or should, be calibrated and whether the 'reduced SHE model' with a limited number of free parameters and spatially uniform parameter values of large regions actually is more physically based than more conceptual models would be. This means also that the parameters might not be that 'physical' after all and the unrealistic root depth could also be an effect of compensating other structural model errors.
As interesting as such a discussion could be, I don't think it is needed here in its full length, but I would recommend to mention these issues at least and to refer to previous discussions on these issues (e.g. the classic Refsgaard-Beven discussion).

This is a good comment and an issue that could be discussed much more. We have tried to elaborate on this in the revised manuscript by including reference to Refsgaard and

Beven and by highlighting the issues of physical realism in model parameter values applied at this scale.

We also feel we should point out, that the parameters are not spatially uniform in the current model setup. There is a high degree of spatial heterogeneity of most input variables. E.g. the geological model is a complex 3D model consisting of up to 11 computational layers and 10-15 geological units. However, for the calibration the parameter values of some of these units are tied to each other based on their sensitivity and similarity. Likewise the unsaturated zone and land use maps are quite detailed and based on large datasets, their spatial distribution or the relation between parameters for the different land use classes is however not changes during calibration, only the absolute level. Two parameters can however be regarded as spatially uniform (although they vary for each model domain), that is the drainage constant and the leakage coefficient. We agree that this is a rather crude assumption; however previous efforts to distribute these parameters spatially have not been fruitful. Thus, due to its basis on comprehensive spatial physical data we consider the model, in spite of the calibration of a few parameters, to be more physically based than traditional lumped conceptual rainfall-runoff models.

The approach, where distributed physically based models are calibrated based on a limited number of free parameters while maintaining a high degree of spatial variability is believed to be sound and realistic, and it is a standard procedure that has been tested in many studies (e.g. (Andersen et al., 2001; Henriksen et al., 2003; Refsgaard, 2000))

Please also discuss how your results potentially might be affected by the type of model you were using. Would you expect the results to be similar or different (in which aspects) if you had been using a more conceptual or lumped model approach ?

A more conceptual or lumped model approach might have given similar results regarding model performance.
However, using a much simpler model with non-physical parameters, it would have been very difficult to evaluate which model parameterisation was more favourable. E.g. In our study the winter precipitation bias is compensated by increased summer evapotranspiration, due to the very low potential ET in winter. This will "twist" the optimized parameters in an unrealistic way, which can be evaluated through the optimal parameter values of e.g root depth. Of course, even the root depth will be an effective parameter at this scale, but it still has to have some similarity to realistic observed values. In contrast some more conceptual models will have tuning parameters that inhibit detailed analysis of model input biases, e.g. the VIC model has a precipitation correction parameter, which in itself could give valuable information on model input error, but in combination with other calibration parameters will disturb the use of parameter values in the model evaluation.

In addition, the fact that the model optimizations give very similar parameter values across all the independent model domains gives us a strong indication that the optimized parameters and especially the differences found between the tested precipitation inputs are robust and significant.

We have elaborated on these issues in section 3 and in the discussion of the revised manuscript.

4) The focus in the modelling is much on subsurface processes. One could argue that above-surface processes are more directly related to the way precipitation data has been corrected. I would expect that interception parameters would also be found to be related to the correction method, if they had been calibrated. Keeping the subsurface parameters fixed and calibrating parameters related to the vegetation, thus, would be a valuable additional test.

The surface parameters are represented mainly through the root depth, which is the main control on evapotranspiration. But also the leakage coefficient and the drainage constant are important parameters for surface water. Interception could have been included in the calibration. However, interception is not considered to have a major impact on the results because of two combined factors: First of all leaf area indexes are very small during winter in Denmark because 90 % of the land area is agriculture with bare soil in winter. At the same time the differences in precipitation between the two inputs tested in this study lie entirely during winter, therefore changes in interception parameters will not cause major differences in runoff.

5) The NSE values are quite low for a number of catchments (Fig 8). Especially when one considers that the model has been calibrated NSE values below 0.5 seem to indicate some problems with the data. Could you comment on this ? Would it be reasonable to exclude catchments, where the calibrated model gives low NSE values from the further analyses ?

This is a good point. These stations could have been excluded from the calibration, but at least they are included in all calibrations independent of precipitation input, and will therefore have a similar effect on the results for all optimizations. The thoughts behind leaving poorly performing stations in the model are elaborated below.

One could argue that the worst performing catchments <u>are</u> actually excluded from the result analysis, because of the way the results are displayed. The NSE figures (e.g. Fig 8) where stations are sorted according to their performance enables the reader to compare the different models by comparing exceedence values and put less emphasis on the very low tail of the stations. Also the reference to the figures in the text addressed the percentage of stations above or below specific thresholds, not the actual NSE of individual stations or arithmetic means of all stations.

There are as mentioned, very probably some issues with some of the stations. Some discharge stations represents very small catchments, and these are especially vulnerable to any non-natural disturbances which are not included in the model. This could e.g. be groundwater abstraction for small fishfarms which is lead directly to the stream, and which will significantly impact the baseflow of the stream. Or small structures like weirs or dams on streams, which will affect the hydrograph significantly. Such issues have been addressed through thorough analysis of existing databases and maps and some stations have been removed from the analysis for this reason, but in cases where no evidence has been found stations are not removed. In other cases, the simplicity of the model and errors in the underlying geological model causes the model to perform poorly on specific small streams.

Analyses

6) There is a minor difference in the total amounts of precipitation in the two correction methods and this is used to explain the differences in model performance and parameter values. However, I would think that the better temporal distribution is at least as important, if not much more. I would therefore like to suggest another test which might provide useful information on this issue. To focus on the temporal aspects you could scale the constant correction in a way so that the total precip amount corresponds to the 'dynamical' corrected one on (1) an annual basis or (2) long-term mean monthly (seasonal) values. Using these two constant, but scaled corrected precip series you could distinguish between the effects of the general differences in precip (water balance effects) and the effects of the temporal distribution.

This is a very good point, and we had considered that, but decided it might disturb the overview of the paper.
However, since the reviewer requests this analysis we have made some further testing along the lines suggested.

We have prepared a third precipitation dataset corresponding to suggestion (2). New fixed nation wide monthly correction factors for the calibration/validation period (2000-2007), hereafter named Current Mean Monthly correction (CMM) (see reply below for explanation of new abbreviations). We believe this is the best way to address the issues of temporal distribution and have added an extra analysis based on this input.

The analysis is carried out for only two model domains (1 and 5, which are large and contain the most discharge stations (93), Calibration times of app. two weeks pr. domain inhibit us from a national assessment.
For the two model domains we have redone the calibration using the new precipitation input, which has new monthly mean correction factors, but no daily or spatial variability in the correction factors.

We still believe that the initial analysis is the main analysis of the paper, because in reality the modeller's choice will be between "old" standard monthly correction factors (HMM) or daily dynamic corrections (TSV), because in order to calculate new monthly factors one needs to calculate the daily, and hence it makes most sense to use those directly. Having said that, the additional analysis is still very interesting regarding the separation between the effect of removing mean bias and introducing higher temporal and spatial variability in correction factors.

Results of additional analysis:

The results of the additional analysis are that the main difference between the model results using the Dynamic (TSV) and Standard (HMM) precipitation during calibration is in the removal in the mean bias, not in the improved temporal and spatial resolution of correction factors. This is seen because the model with the new precipitation input, with average updated corrections (CMM) performs very similar to the Dynamic (TSV) for the calibration period (NSE and WBE). See top figures below (they will be included in the revised manuscript).

It should however be mentioned that the calibration period 2000-2004 has a strong tendency to a systematic bias (Fig 4), meaning that on a seasonal basis, the difference between the dynamic (TSV) and the New mean dynamic (CMM) are quite small. In addition, the stream discharge of especially Domain 5 is largely groundwater controlled, causing the sensitivity of the very fine (daily) temporal resolution to be smaller.

When analysing the results for the validation period, the two split sample years 05/06 and 06/07 are used to illustrate the impact of correction method for validation years with different biases. As illustrated in Fig. 4 in the manuscript, the year 05/06 has a TSV correction below the HMM correction whereas the year 06/07 has higher corrections more comparable to the general bias in the calibration period. It is evident from the results from the two validation periods, that when biases are systematic, meaning that there is a general tendency to over or under estimation of precipitation the aggregated CMM correction will produce results quite similar to the TSV correction (Validation year 06/07), especially for NSE. However, when biases are of alternating direction compared to the calibration period, such as for the year 05/06, then the CMM correction yields results more similar to the HMM correction, indicating that there is an improvement by using the TSV correction. Unfortunately our data did not include shifting annual biases during the calibration period, which could have given other results in favour of the TSV correction compared to the CMM correction also for the calibration period.

Other results might also have be seen in catchments which are controlled by much faster rainfall-runoff processes, such as overland flow, which is almost negligible in Denmark. And, if the Dynamic (TSV) correction method had been based on more local input data, such as local wind speed, temperature, intensity and precipitation type, larger temporal/spatial variation might have been observed in the TSV correction compared to the CMM, which might have emphasised the detailed temporal aspects, which can be "smoothed" out a bit by the use of off-site data.
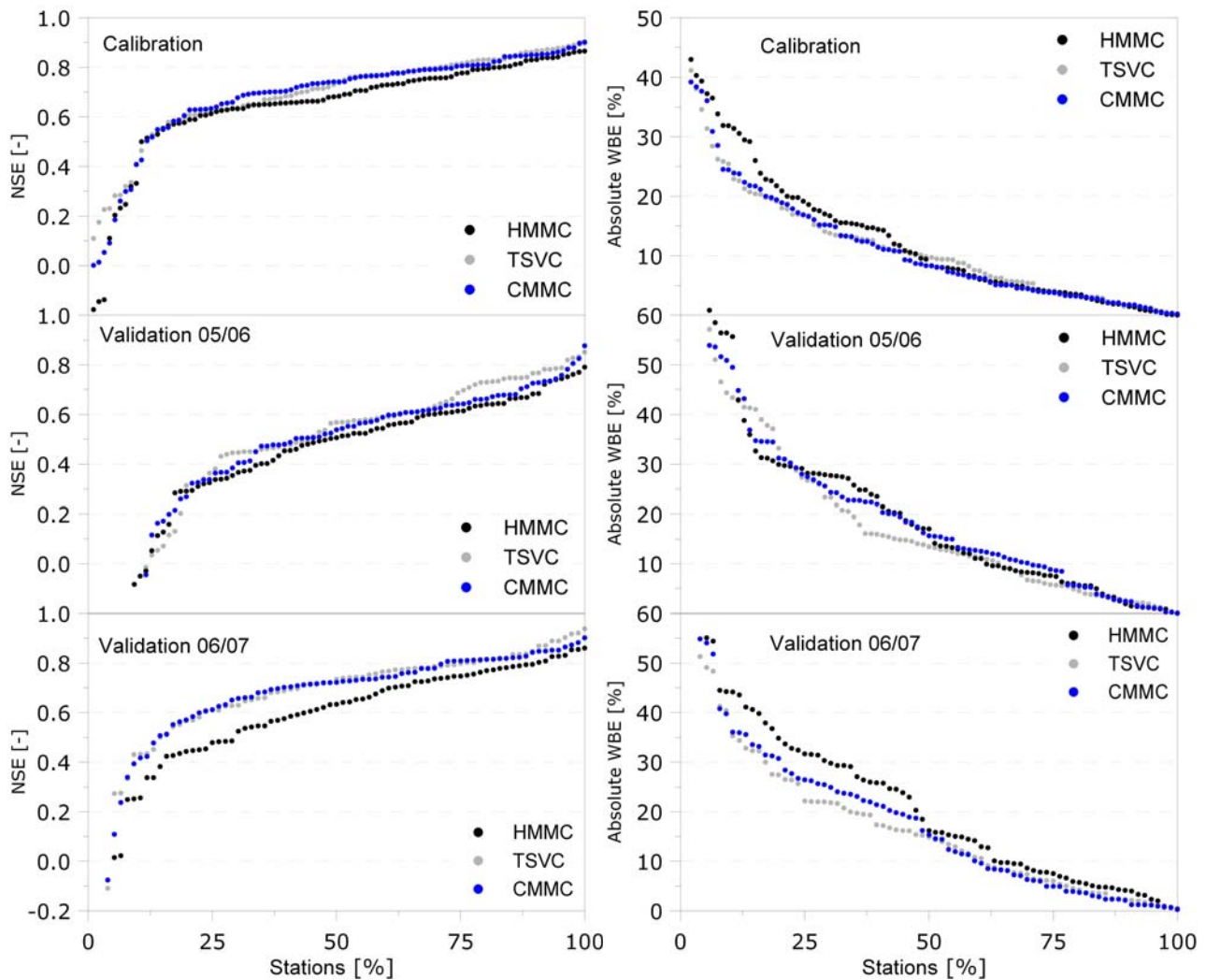
Figure 19: Calibration and split sample validation results for discharge stations in model domains 1 and 5 including the current mean monthly (CMM) corrected precipitation data.

Further comments:

Title: as there is a large focus on solid precipitation I recommend to change rain-gauge to precipitation-gauge (also in the text the terms rain and precipitation should be used more consequently)

Good idea, corrected throughout the revised manuscript

Instead of the term 'dynamic correction factors' something like 'time-variant correction factors' would be better, not sure what is meant by dynamic (more than variable).
Also the term 'dynamic precipitation' in figure 3 is a bit vague, please always be clear that you refer to the correction method.

Good point. The issue is that the "dynamic" correction is variable in both time and space. In order to comply with the reviewer we have changes the naming: Standard -> Historic Mean Monthly (HMM) correction and Dynamic -> Time and Space Variable (TSV) correction, the new averaged dynamic -> Current Mean Monthly (CMM) correction. This change has been applied throughout the paper, including figures and tables

Avoid variable names like CF, which might be understood as C times F

Changed to C.

As it is now, both results and discussion is mixed in the results section. I would recommend
to separate the two clearly into two sections.

Good point, that is included in the revised manuscript where separate section (4, 5, 6) deals with results, discussion and conclusion

The issue of precipitation correction obviously has been addressed for a long time (although not so much more recently), and it might be appropriate to refer some more of this work such as the detailed studies by Boris Sevruk.

Good idea, done in the introduction

Anonymous Referee #2

Received and published: 7 May 2012
REVIEW – Manuscript: Stisen et al., 2012, HESS

On the importance of appropriate rain-gauge catch correction for hydrological modeling at mid to high latitudes

Overall Review
The manuscript presents an analysis of how applying precipitation correction factors affects simulations of hydrological dynamics. The authors applied two specific correction methodologies suited for Scandinavian regions: (i) mean monthly correction factors (standard correction) and (ii) dynamic factors (time evolving and daily time step). These corrections were applied to the rainfall fields used to force a hydrological model. Better performances and more plausible values of model parameters were obtained in the case of a dynamic correction. With this study the authors highlighted the importance of having "a correct" input of precipitation, and that automatic calibration of hydrological model are very prone to lead to unrealistic parameters just to compensate for wrong inputs. Despite the fact that the paper is well written and clear, and the results are well presented, I have major concerns about the novelty and generality of the presented study (see below).

MAJOR COMMENTS
1) The introduction is lacking a broader picture of the consequences and importance of

the study. How the results obtained for a specific correction method and for a particular geographic area can be of general interest? How other correction techniques compare with the ones used by the authors? A wider discussion referring also to previous literature and stressing the importance of having a, as precise as possible, meteorological input, rather than a very detailed calibration procedure might make the paper of interest to a larger audience. This discussion is only partially presented in the conclusion but it might be highly significant for most of the reader that are less interested in the specific Danish case.

We acknowledge the comment of the reviewer, and have broadened the abstract and introduction and discussion in the revised manuscript. This includes more focus on the generic perspectives of the study and its implications for climate change impact assessment. Also we have included a section in the introduction on previous literature on more general aspects of precipitation bias and uncertainty vs. model calibration.

We believe our results are of general interest because they highlight the importance of updated/dynamic correction factors through a robust modelling exercise. Any modelling application in regions where solid precipitation is involved will be facing the choice of precipitation correction method and the dilemma between input bias and model calibration.

Correction methods are typically empirical and local/regional, and dependent on the dominating precipitation type, wind direction etc. Other countries/agencies will have different but similar correction methods, basically all based on temperature, wind speed, intensity and precipitation type. So independent on the exact correction method, our analysis illustrates the importance of calibrating a hydrological model based on the best available precipitation input, rather than adjusting biases in precipitation through calibration. This has been highlighted in sections 1 and 2 of the revised manuscript.

2) Another major comment regards the novelty of the presented study. The fact that the methodology presented by the authors has been already evaluated at the catchment scale (Page 3610. Line 26-28, see Stisen et al., 2011b) in Denmark and that the main contribution of this study is to extend it at the national level is of great concern to me. The Stisen et al., 2011 (VZJ) paper has a very similar structure to the paper submitted to HESS. They use the same precipitation correction methodologies (standard and dynamics), the same calibration procedure of the hydrological model, and they have very similar conclusions. Now, I'm not fully sure the extension of the analysis from one catchment to the entire Denmark is enough for considering this study as "new".

The previous paper in VZJ showed some interesting results for a particular very sandy and groundwater controlled catchment. This gave us a hypothesis about the importance of dynamic precipitation correction, that we feel should be tested on a much larger scale and applied to hydrologically different regions, in order to be of more generic interest. And actually East and West Denmark are significantly different, regarding climate, geology and hydrological responds. While western Denmark, where the previous study including a single river catchment was performed, is very wet, sandy and has a slow hydraulic responds, Eastern Denmark is drier, the geology is dominated by moraine clay and the

hydraulic responds is much faster. In addition, the difference in precipitation between the correction methods is spatially very different (Figure 1), which is not reflected in a single catchment analysis, where both correction methods are basically spatially uniform. The current study, addresses a combination of temporal and spatial improvements in precipitation correction, which was not possible in the previous study due to the limited size of the study area. Our revised manuscript tries to highlight the spatial aspect of the study and the more general conclusions that can be drawn from a wider application.

Ideally, a more comprehensive analysis could have included other regions in the high latitudes. However, we feel that going from a single catchment in the previous study to the entire country (43,000 km^2) in the present study with the large variations in hydrological regimes outlined above is sufficient to derive clear and robust results, which are of general interest to hydrological modellers working in cold regions. An extension of the study to other countries would have included another uncertainty, as the data base (equipment/gauges at climate/precipitation stations, methodologies behind gridded climate data products, etc) would be different.

We believe that the revised paper, in addition to testing the hypothesis by using the national model, includes a range of further analysis and has much more focus on the generic implications (see also reply above) and spatial aspects of this study. We have tried to emphasise these novel aspects in the Discussion and Conclusion of the revised manuscript.

MINOR COMMENTS
Section 1.
Page 3609. Line 8. I would suggest the author to refer also to the work of Nespor and Sevruk (1999) that extensively discuss the problem of wind induced undercatch.

Good idea, corrected

Page 3610. Line 16. I would suggest the author to revise the sentence because recent literature (Ryu et al., 2011) suggests that relatively high-resolution estimates of regional and national evapotranspiration might indeed become available in the near future.

We follow the developments in remote sensing of evapotranspiration closely and tend to disagree with this comment, because even with the recent advances in remote sensing based estimates of evapotranspiration, these estimates are still basically just models and associated with great uncertainty, typically in the order of 25 % on estimated ET, which makes is very difficult to utilize as a comprehensive calibration target. As for the paper referred to by the reviewer (Ryu et al., 2011), the presented estimates of ET has an RMSE of app. 30% even for Annual ET at the large Basin scale, which highly limits its use for calibrating the water balance for individual subcatchments on a daily time step.

Section 2.
Page 3612. Line 2. I would invite the author to separate lambda_solid from lambda_liquid as done in the equations.

Good point, corrected

Page 3613. Line 3-15. The authors made several assumptions for applying the dynamic correction of precipitation, especially as far as concern wind speed. I understand that this is unavoidable for obtaining final results on the basis of available data, however, could the author provide some sensitivity or test on how these assumptions are affecting the correction factor. The authors quote other papers (Vejen 2005; Allerup et al., 2000) but I suppose that it would be possible to provide more explanations on the reasons why these assumptions are justifiable.

Based on this recommendation and the comment from the other reviewer we have expanded the section describing the assumptions behind the dynamic correction. Please see, the revised manuscript. We do however not believe that a detailed sensitivity study would be within the scope of this study, because this has been described in detail in the referenced study (Allerup et al., 2000). They have studied the effect of using off-site information with the same correction model and for Danish conditions.

Section 3.
I would state upfront that all the simulations and modeling are done at the daily time scale.

Good idea, done

Page 3616 Line 5. I think that more than "large model setups" should be "simulations of large areas", I'm not sure what a large setup is.

Done

Page 3616. Line 19. I think "contract" should be "contrast".

Done

Page 3617. Line 5-17. Maybe a map of the geological units used in the model will help the reader.

The 3D geology is quite complicated and very different from region to region. Since the geology is not just a 2D spatial map, the geology should be visualized by several geological cross sections, with individual dominating geological units. We believe this is not of major importance to this paper, especially since there are already many figures.

Reference
Nespor and Sevruk 1999, Estimation of Wind-Induced Error of Rainfall Gauge Measurements
Using a Numerical Simulation. Journal of Atmospheric and oceanic technology 16 450-464.

Ryu et al., 2011. Integration of MODIS land and atmosphere products with a coupledprocess
model to estimate gross primary productivity and evapotranspiration from
1 km to global scales GLOBAL BIOGEOCHEMICAL CYCLES, VOL. 25, GB4017,
doi:10.1029/2011GB004053.

References

Allerup, P., Madsen, H., and Vejen, F.: Correction of precipitation based on off-site weather information, Atmospheric Research, 53 (4), 231-250, 2000.

Andersen, J., Refsgaard, J.C., and Jensen, K.H.: Distributed hydrological modelling of the Senegal River Basin - model construction and validation, Journal of Hydrology, 247 (3-4), 200-214, 2001.

Henriksen, H.J., Troldborg, L., Nyegaard, P., Sonnenborg, T.O., Refsgaard, J.C., and Madsen, B.: Methodology for construction, calibration and validation of a national hydrological model for Denmark, Journal of Hydrology, 280 (1-4), 52-71, 2003.

Refsgaard, J.C.: Towards a formal approach to calibration and validation of models using spatial data, In: Grayson, R. B. and Blöschl, G. (Eds.) Spatial patterns in catchment hydrology: observations and modelling, Cambridge University Press, 329-354, 2000.