GENERAL COMMENTS

In their paper „Statistical downscaling of precipitation: state-of the art and application of bayesian multi-model approach for uncertainty assessment" M. Z. Hashmi, A. Y. Shamseldin and B. W. Melville downscale GCM precipitation projections for the region of the Clutha watershed in New Zealand. This is done by combining results of three downscaling techniques by means of a Bayesian approach (cf. Tebaldi et al. 2005). The authors also claim to perform an uncertainty assessment of the three downscaling methods applied. However, the referee finds a mismatch between the announcements in title and abstract and the content of the paper itself. Therefore I recommend a profound reorganization of the paper. In case this is done, the paper might lie within the scope of Hydrology and Earth System Science by being an applied research paper.

Method:

The Bayesian approach presented in Sections 4.1-4.14 is completely identical to the approach presented in Tebaldi et al., 2005. The authors furthermore claim to present a novel downscaling methodology, namely gene expression programming (GEP). However, the description of this methodology in Sec. 5.3 remains shallow, forces the reader to accept the GEP approach as black box and does not reveal the authors proper proportion of the development of this methodology. If the authors want to claim to have parts with methodological development in their paper, I recommend an explicit description of the steps done for the downscaling with GEP. Furthermore, the authors treat the output of the stochastic downscaling procedures as point predictions and do not use the potential of, e.g., weather generators, to deliver uncertainty bands together with the predictions. I propose to include the uncertainty information given by the single stochastic downscaling models in the Bayesian model merging strategy.

Uncertainty assessment:

Furthermore, in my view the uncertainty assessment announced is not done in the paper. Sections 3.1. or 4.1.5, which should contain this assessment, lacks a general classification of the uncertainty assessed by the authors which gives the false impression that here a universal approach is presented. Even if the authors decide to skip a section of a general overview of uncertainty assessment strategies for stochastic downscaling methods, they should clarify the assumptions and therefore the limits of applicability of the approach they present: The authors find for the calibration period a varying performance of the three downscaling methods used, that is no best method can be identified. Therefore they combine the downscaling output by means of a Bayesian approach (cf. Tebaldi et al, 2005). So far so good, but I disagree with their conclusion that Bayesian multi-model combination always reduces uncertainty, this has to be checked for every application, which the authors did not do for their application (see below). I therefore recommend to profoundly rework the uncertainty assessment parts in the paper and to adapt the conclusions to the content of the paper.

Application:

The paper presents a nice application of downscaling of precipitation for the region of the Clutha watershed in New Zealand. This could be the strong part of the paper and could make it suitable for the HESS journal as applied research. However, the referee misses important parts of the analysis of the application. I think the authors should verify their assumption that the estimated weights for the models in the calibration period are transferable to a prediction period for their application by, e.g. comparing the WMME predictions to observations in the verification period. Furthermore the authors should compare the WMME projections with the projections of the output of each single downscaling methodology to underpin the statement that the combination of models reduces uncertainty. Also, the authors aggregate the data in several levels: they aggregate over the whole prediction period of 30 years and the whole Clutha watershed of about 150 x 300 km. Therefore I recommend as well a comparison of the WMME results with the GCM predictions to show that downscaling advantages nevertheless remain. Finally, I miss an interpretation of the projection results, i.e. the consequences of the precipitation changes for the Clutha watershed.

Miscellaneous:
I found the style of the paper, the exactness of expressions used, and the English partly very poor and the paper should be revised regarding these issues.

SPECIFIC COMMENTS

Scientific Quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?

- abstract p6536 line 11 and introduction p 6537 line 11: statements about uncertainty analysis missleading: The paper does not adress „the uncertainty analysis“ associated with statistical downscaling of watershed precipitation. It has a very narrow aspect of uncertainty analysis, namely the models are weighted according to their bias and precision.
- abstract p 6536, line17: what does „efficient“ mean in this context? The Bayesian method combines the model output regarding some criteria (bias and variance in the calibration period). It has not been shown that these weights can be transferred to the prediction period.
- p 6538, line 7: „The third model is an artificial intelligence data driven model developed by the authors using the Gene Expression Programming (GEP) to create symbolic downscaling functions.“ If this is the case, please outline in section 5.3 what kind of model you developed, i.e. which links were chosen, characteristics of the model, etc, etc.
- p 6538, line 23: „Secondly, a discussion about how to deal with downscaling uncertainty using multi-model ensembles is provided“. I cannot find this discussion. In Sec. 4.1 just outline the setting of the Bayesian WMME approach.
- p 6540, lines 19-21: „Being subjective in nature, this approach is less ..... and hence was not used in this study in developing the multi-model downscaling ensemble“. The referee advises the author to skip this sentence or to support this statement in the analysis, i.e. to compare of the performance of the Bayesian WMME with and without the method of Ruosteenoja et al. (2007).
- p 6542, line 8: „Due to many noted reasons discussed in the introduction of this paper, the results obtained from downscaling models may have a considerable amount of uncertainty and .... will be wrong“. The referee does not find a detailed discussion about uncertainty in the introduction. The referee rather suggests to extend this discussion in section 3.1. Up to now this sentence discredits all deterministic downscaling methods without giving any justification. Up to now I regard the discussion presented in this paper as being too shallow and recommed revision regarding this issue, see, e.g., Fowler et al. (2007).
- p 6542, line 19 ff: „Although a simple average approach has shown definite advantages over a single model approach in terms of robust uncertainty assessment (Hagedorn et al., 2005)“. The referee advises to reformulate this sentence and to make the statement less strong. Model ensemble averaging for example will worsen the results of „the best model“, in case this model always outperforms the other ones. Furthermore, the term „robust uncertainty assessment“ has to be precised - this can mean a lot of things depending on the context.
- p 6544, lines 16, „λ_0 is the observed variability of moean monthly Clutha precipitation as given .... future Clutha precipitation“. Please outline the relation between X_0 and μ, wich gets not clear so far.
- p. 6545, Sec. 4.1.2: all parameter priors have been chosen as uniformative. This contradicts the in the introduction stated advantage of Bayesian analysis to include expert knowledge in the uncertainty assessment. The referee suggests to discuss this aspect in the conclusion.
- p 6545, lines 17-19: The referee suggests to outline here that in the approach used in this paper (Tebaldi et al., 2005), the data is seen as observations and model output, i.e. y=(X_0, X_1,

X_2, X_3, Y_1, Y_2, Y_3). This is rather unconventional and thus not clear ad hoc.

- p. 6546, lines 8-14: This paragraph is lacking the reference to some assumptions: (Y_i-v) is only a measure for „convergence“ in case one assumes that the models vary (with a normal distribution) around the model mean, which is assumed to be near the „true value“. Outliers of the model ensemble are therefore punished with less weight. This assumption may not always hold.
- p 6546, lines 16-21 (last paragraph). Please state here which method you used for MCMC simulation and give a reference.
- Section 4.1.5. Uncertainty assessment: The referee misses a detailed classification of the type of uncertainty assessed (with references, etc). In the introduction it is stated that the uncertainty of stochastic models shall be adressed. But then only the result is presented, that the three stochastic downscaling methods have a different precision in the calibration period (results summarized in table 5). I expect in this section a detailed outline of the assumption of the weights used and reference to potential other uncertainty assessment methods. Furthermore the consequences of the assumptions should be outlined. For example, the WMME only uses statistical characteristics as weighting criteria (bias and variance in a calibration period), no process oriented criteria. Correlation of the stations is not regarded at, and so forth. What consequences for the results does this weighting criteria selection have and to what use this weighting is limited.
- Section 4.1.5, Uncertainty assessment: Parts of this section are missplaced. What has the definition of the change detection used by the authors to do with uncertainty assessment, and so forth.
- p 6544, line 10: „.... for the month represented by the data“. Here it is stated for the first time that the whole analysis framework is applied separately to the data of each month. Please make this clear beforehead, for example in Fig. 2.
- p 6548, line 21ff: „....that the correlation values obtained are well below the acceptable limit as indicated in previous studies (e.g. Hessami et al. 2008)“. Please outline what is the acceptable limit and why/in which circumstances.
- p 6549, lines 13ff: „The predictor selection process is consistent with .... The 10 chosen predictors were used for calibration of the downscaling model“. 10 predictors for 30 datapoints in the calibration period seem a lot to me. In case the selection process did not include a step where the complexity of the regression model used (i.e. number of predictors chosen) was counterchecked with the amount of variance explained, e.g. an ANOVA criterion or so, the reviewer suggests to include such a step.
- p 6549, lines 23 ff. Please give references for the characteristica checked and/or a short explanation, e.g. „variance inflation“, „bias correction“.
- p 6550, line 12: The text suggests that more characteristics than monthly mean and monthly sd of precipitation have been compared. What else has been used and what are the results?
- p 6551, line 3: „As discussed in Sect. 3, studies have shown ...“ These studies have been refered to, there was no discussion of this issue in the sense of outlining pros and cons of these methods, applicability, etc. Thus either reformulate your statement here or discuss this issue.
- Section 5.3: GEP model. Here the referee is missing a detailed outline of the procedure. The class of potential link functions, the symbol selection process and the rest of the setting (modeling of the noise, etc.) is not described. Furthermore the results in Fig. 10 suggest some kind of overfitting (underestimation of the variance, e.g.). So how complex is the regression model used?
- p 6552, line 14: „and then the relative weight (ω_i) to be “. The referee suggests to define this relative weight in a formula and describe its meaning. It is the first time that ω_i is mentionned.

- Sections 6.1-6.3. Description of the results: here seems to be a confusion between description of the downscaled values and the GCM. At p 6554, line 6, for example, the authors refer to GCM projections but describe Fig. 9, which shows averaged LARS-WG output. Please streamline the text accordingly.
- p 6555, lines 17-19: “In this way, the MME has taken into consideration the strength and the weakness of each model and produced a downscaled output which would be more reliable than either of the individual models.“ Here the reviewer disagrees. The MME resulted in weights for the models according to their performance in the calibration period. Then, under assumption that bias and precision of the models stay the same for the prediction period, these weights have been transfered to the prediction period. The reliability of this assumption for the application given has not been checked by the authors, e.g., for the verification period. It has hot been verified that the multi-model ensemble mean ν does lie nearer to the accordingly averaged observations or the accordingly averaged output of each of the three stochastic downscaling models. Even if so, it is not clear that this assumption holds for the prediction time period 2070-2099. Therefore the referee asks the authors to broaden their analysis and to reformulate their achievements. Furthermore, the referee considers „bias and precision“ not to be the same as „strength and weakness“ of a stochastic downscaling model and again asks the authors to reformulate their statements accordingly.
- p 6556, line 13 „Three well reputed downscaling models namely SDSM, LARS-WG and GEP were used.“ This is contradictory to the former claim of the authors to have used GEP for the first time in a downscaling context, so please harmonize the according passages.
- p 6548, Section 5.1 and Fig. 3: Please define the „maximum range of correlation“ and give a reference.
- Section 4: The authors present a Bayesian method to combine output of three stochastic downscaling procedures. Here the authors treat the output of the stochastic downscaling procedures as point predictions and thus do not use the potential of, e.g., weather generators, to deliver uncertainty bands together with the predictions. One could include the additional information of the variability of the stochastic downscaling models for example by a different estimation of the precisions λ_i. The referee proposes to include an according study in the analysis.
- Fig. 13: I miss the interpretation of the results of Fig. 13. What does this mean for the Clutha watershed? Or, for example, assessment of variability: is the distribution of percentage change broader or narrower than $P(Y_i|\Theta)$, i.e. comparison of WMME and single downscaling model projections.

Presentation Quality: Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?

- Repetitions occur in the text, for example:
  p 6539, lines 13-15 and lines 22-25: it is referred two times to the HadCM3 output used, once a time period of 1961-1989 and the other time a time period of 1961-2000 is mentioned. I recommend to streamline the paper regarding this issue.
- p 6537, line 26: „There is very limited research regarding uncertainty analysis associated with statistical downscaling...“. Please name some references, e.g., Fowler et al. 2007.
- p 6538, last paragraph: please refer to the according chapters/sections.
- p 6542, lines 12 ff: „To our present knowledge, there are only limited studies which deal with the uncertainty analysis of downscaling results and the first attempt .... is made by Khan et al. (2006).“ The referee suggest to reformulate this sentence. Just to give one example: STARDEX is an EU project, which has done intercomparison studies for

deterministic and stochastic downscaling methods for extremes and has been run from 2002 to 2005. Certainly there are papers published before 2006, which treat uncertainty of stochastic downscaling methods.

- p 6547, lines 21 ff: „TME has been implemented in a computer program developed using the statistics package R which can be downloaded ....“ Please name the software package and the authors of the package. Move the whole reference to the software to another section, for example 4.1 or the appendix.
- p 6548, line 14. Clarify the term „Ophir2“.
- p 6551, line 23: „The results obtained using GEP show .....“. Please name the figure to which this sentence refers to.
- p 6552, line 7. „A number of initial samples were discarded ....“. How many? Furthermore there seems to be a typo in the text: If the authors took 5000 samples and saved every 50th iteration, this would leave 500.000 iterations for the burn-in period to obtain the 750.000 iterations mentioned.
- Figs 7, 9, and 11 represent approximately the same information, obtained with different downscaling methods. Therefore the referee suggests an aggregation of the three pictures to be able to compare the results. Additionally a comparison with the WMME results would be interesting.
- **Precision of text**:
  The precision of the text is inadequate at some text passages. Either the text is too spongy or terms are created without specification of their meaning, thus leaving too much room for interpretation. Some examples are:
  p 6536 line 17: „ensemble strategy“
  p 6539, line 6: „long term annual mean flow“
  p 6538, line 7: „artificial intelligence data driven model“
  p 6538, line 12: „ensembling information“
  p 6537, line 14 „working principles involved in the operation of the technique“
  p 6550, line 5 „long term daily information of the climatic parameter“
  p 6551, line 15 “powerful soft computing package“
  p 6555, lines 24-25. „Examining Fig. 13 in terms of IQR as a measure of uncertainty, a variable trend can be seen of monthly ....“. The paragraph is unclear, what is meant with „variable trend“?
- **English language:**
  I am no native speaker, but I have the impression that the English language (expressions and grammar) should be revised, especially the abstract, the introduction chapter and the conclusion chapter. Here some examples, the list is not complete:
  p 6536, line 10 ff: This paper addresses the uncertainty analysis associated with statistical downscaling of a watershed precipitation (....) using results .....
  p 6537, line 15 ff: Although the statistical downscaling is very popular and extensively used in many studies (Christensen et al., 2007), it usually performs well only for the conditions and regions where it was originally developped.
  p. 6549, line 8: typo: NCEP instead of NECP.
  p 6556, line 18 ff: The large scale data of HadCM3 model has been used for baseline period and future period .....
  p 6550, line 20:“....based on SRES A2 scenario run are used ...“ „one“ or „a“ is missing.
  p 6550, lines 1-4: sentence structure not adequate.
  p 6550, lines 1-6: structure of text is inadequate. There is doubling of information, for example regarding the capacity of LARS-WG to synthezise data.
  p 6551, line 15: „a powerful soft computing package“. The referee considers this style as not being appropriate for a scientific paper. It rather resembles a software advertisement.

p 6555, line 5: „Fig. 12a and b is pictorial representation ....“ Expression.
p 6556 lines 17-20: sentence structure, some a's missing.

References

Fowler H. J., S. Blenkinsop, and C. Tebaldi. Linking climate change modelling to impact studies: Recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology* , **27**:1547–1578, 2007. doi: 10.1002/joc.1556.

Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles. *Journal of Climate*, **18** (10), 1524–1540, 2005.