Editor Decision: Publish subject to minor revisions (further review by Editor) (05 Jul 2017) by Thorsten Wagener

We thank the reviewers and Editor for taking the time to review our manuscript again. Your comments have definitely improved the clarity of the paper. Editor/reviewer comments are black. Responses to comments by the authors are blue. Note that line references refer to the most recent version of the paper.

Comments to the Author:

The reviewers have gone through the revised manuscript again, and, while they accept that things have improved, they still have quite a few comments about how to make the manuscript more accessible and clearer. I selected minor revisions for now. However, I still expect you to address all the points that I list for bother reviewers in detail. All these comments should make the manuscript easier to understand and therefore will help you. So please take these comments seriously. I will send the newly revised manuscript out to reviewers again if I do not feel that you have done so. There are 7 comments by reviewer 1 and 3 main comments by reviewer 2, with some additional minor ones.

We think reviewer #1's comments 2, 3, 4, 6 and 7 reflect a misunderstanding of the aim of the paper. We've tried to explain this in more detail below and have modified the paper to try to prevent such misunderstanding.

Report #1

This paper should be revised significantly before publication. Some information is missing, such as an equation showing how model results are differenced with observed data. There are typos and some phrasing indicating hurried preparation. My comments are as follows:

Scientific comments:

1) Line 5, pg 9; use of equation 1 relates x1 to x3. Constraining one parameter constrains the other. With a line source (line 18, pg 8) this is presumably distributed uniformly from surface to top of plume.

Yes, your interpretation is correct. This is noted on P8, L25 and P9, L5. We have changed "evenly" to "uniformly" for clarity.

2) There are 18 parameters to constrain, and only the distribution of cloud load with time to constrain them. Perhaps the scheme acts to just confine the fit to the data manifold, and so impotent parameters are ignored. But the number of parameters does seem superfluous. For example, do the authors need two turbulence parameters and two loss parameters, when they only have scalar transport in a given wind field?

The aim of this study is to better understand the influence of source and internal model and parameters on the simulator (NAME) output (horizontal distribution of ash column loading). To do this we need to include simulator parameters, in this case this includes two turbulence parameters and two loss parameters. Note that similar emulation techniques have been successfully applied to a variety of physical problems in the literature.

We emphasise that our aim is not to *constrain* parameters (e.g. using actual atmospheric observations of the ash cloud), but to build an emulator to replicate the model behaviour. With enough runs of the model there is no problem in principle of understanding the influence of any number of parameters on column load. We've tried to make clear the difference between constraining parameters with real observations and building an emulator to approximate the model (see p4 line 27 to p5 line 3, p5 lines 21-25, and p17 lines 3-4, p19 line 8, p22 line 24 and p24 line 5).

3) Lines 10-17, pg 13; it seems that including the laminar sublayer (very near the earths surface) is a very different scale than the transport of ash in the atmosphere at an altitude of kilometers above the surface. Why include this? By the time ash is near the ground, transport is local and on a scale for which the authors have no data.

We are aiming to understand the influence of as many of the internal simulator parameters as possible so have included this term for completeness. We don't understand the comment about the authors having no data on a local scale. We have no actual atmospheric ash cloud data at all in this study – the aim is to build an emulator to replicate the simulator.

4) Equation 8; this is a linear regression equation with an error u(x), which is then used later in a linear Bayesian analysis. This is fine, but the means by which data are compared to model runs to perform this fit is not mentioned. Do the authors use a quadratic loss function, for example?

In section 5.1 we specify that least-squares regression is used to construct the emulator model in Equation 8 for the fast simulator. The emulator for the slow simulator is fit using a Bayes linear update described in Section 5.3.

We note we are not comparing real atmospheric data with results from (NAME) simulator runs, but are comparing data from runs of the NAME simulator with data produced by the emulator (and, for the fast simulator, are choosing the beta to minimise the sum of the squares of the differences).

5) Pg 21; Again, the authors are using a linear regression, with R squared, but do not explain what differences are used. The form of the equations in the Appendices suggest a quadratic loss function, but the reader should not have to guess.

It has been restated in Section 5.4 and at the start of section 6.2 (and at many other places in the text) that least squares is being used.

6) Lines 10-18, pg 25; the curse of dimensionality needs to be discussed somewhere here. How can so few data (distribution of cloud load with time with a given wind field) be constraining interactions between so many parameters?

Perhaps this comment and comment 6) below under Editorial comments suggest a misunderstanding of the emulator's purpose: it is trying to predict what the NAME simulator will do under given conditions and parameters. It does not, by itself, provide any predictions about real-world behaviour. Emulators are designed to understand what the simulator would do if we ran it at new parameter settings. Given enough runs of the model there is no shortage of data for understanding how the model's column load varies in response to any number of parameters. We have added the following into the introduction: "It is important to understand that emulators are used to model the behaviour of the simulator itself, when parameters are varied. That is, an emulator is designed to predict the output of the simulator under given conditions. The relationship between the simulator output and real-world observations does not have to be considered in order to build an emulator; the ``observations'' used to build the emulator are observations of simulator output, not real-world measurements.”

7) The two most important parameters, plume height and source strength, are tied together by equation 1 and do not vary independently. With any significant wind shear with altitude, there is only a limited range of solutions to be obtained as the line source increases in height, and much of the ash delivered at various heights are spurious sources. The heights are tied together by the nature of the chosen source, and these combined heights spew ash in different directions as the wind varies with height. The different directions are tied together, producing a single solution to be compared with data. With this source, the emulator cannot separate out the effects of different heights, but will use the degrees of freedom provided by the different parameters to try and explain the variations in distribution. A more effective approach would be to consider point sources at specific heights, and allow their strength to vary independently of height. Ash is usually delivered at a narrow range of elevations, despite the large column heights at the source.

We thank you for your comment and suggestion of approach. As stated on P8, L18-22, we have performed emulation procedure using a “thin layer” source (where all the ash is released close to one height) and in this case there is very little difference in the final results on parameter sensitivities. We emphasise that we are not trying to compare with observational data or to explain observed variations in distribution of ash in the real atmosphere. However, if we were doing this, we agree that the height distribution at the source deserves more consideration.

Editorial comments:

1) Line 15, pg 7; 14 parameters mentioned here, 18 in Table 1

This inconsistency has been corrected in the text.

2) Line 15, pg 10; turbulent

This has been corrected.

3) Line 12, pg 12; B is used as a parameter here, and B is used for something different on pg 19.

For clarity and to emphasise the difference between the variables we have changed the font of the B used on P12.

4) Line 20, pg 13; D is used as particle diameter here, and D is used as a vector of simulator runs on pg 19.

For clarity and to emphasise the difference between the variables we have changed the font of the D used on P13.

5) Line 14, pg 17; hosen

This has been corrected.

6) Pg 16; this might be a good place to put exactly how the comparison between model cloud load and observed cloud load is used to construct a vector of differences; is it quadratic? Absolute value? How are simulator outputs compared with data?

See response to scientific comments 4 and 6 above.

Report #2

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

The manuscript has been significantly improved after revision. I think the contribution is interesting and should be considered for publication. However, the manuscript could still be improved in terms of clarity. Below are details of 3 major and several minor suggestions for improvement.

[1] Clarify the connection between emulator development and sensitivity testing of the slow/fast simulators. One may expect the sensitivity testing to be carried out in two stages: first identifying the emulator, then using the emulator as a substitute of the simulator within a computationallyexpensive sensitivity analysis. However after reading the entire manuscript I understand that there is no such a second step, and that the sensitivity testing is a "byproduct" of the emulator identification process itself: the estimated emulator's coefficients (beta) are the measures of the simulator sensitivity to its parameters (x). This could be clarified in the Abstract and Introduction. For example on P. 4 L. 18-19: "This enables the quantification of the impact of each simulator parameter on the prediction of the dispersion of volcanic ash": the sentence could be revised to clarify "how" the quantification is enabled.

On P4 we added "In this study, the impact of the various simulator parameters can be assessed by their coefficients within the emulator. Since the emulator can be evaluated quickly, it can also be used to replace the simulator in any computationally-intensive sensitivity analysis method of choice, though this step is not performed in this study."

On the same topic, the term "active" should be defined in Sec. 6.1 (how is it operationally concluded that a parameter is "active"?) and the link to sensitivity testing established. I understand that the fact that a parameter is active in the emulator implies that the output of the original simulator is sensitive to that parameter, so "finding active parameters" is the sensitivity testing, but this is not stated explicitly. Please clarify.

We have included a definition of active and inactive variables at the start of Section 6.1 and a link to the section of the appendix where the specific details of the process of identification is explained. We also clarified that active variables are not necessarily really impactful: "Note that active variables are not necessarily extremely important parameters; they simply provide some information that would be lost by excluding them."

Identifying inactive/active parameters could be regarded as part of the sensitivity testing, but it's only part, and it leads to a smaller set of parameters for more quantitative analysis. Alternatively one could view it simply as a step in building the emulator. The sensitivity analysis can then be done using the emulator, either just using its beta values as is done here (these can be regarded as zero for inactive parameters, making the sensitivity analysis for these parameters trivial) or by using the full emulator. Our presentation tends to emphasise this second view which we think is clearer.

[2] The description of the emulation methodology (Sec. 5 and associated Appendices) is still confusing. I think the main text still puts relatively too much emphasis on the statistical rationale underlying the emulator and too little on the practical steps for their construction and use. This may

make the manuscript not very accessible to non-statisticians and limit the uptake of the proposed multi-level emulation approach.

We have added in the appendix a step-by-step description of the process used and, for each step, references the section(s) of the paper dealing with the step.

Specifically:

- P. 18 L. 9: "Such an emulator then provides predictions for f (x) at a new x. Since it is a statistical model, this prediction also comes with an associated uncertainty". How are the predictions and associated uncertainty obtained in practice? I guess the prediction is the adjusted expected value of Eq. (9) and the associated uncertainty is expressed by an uncertainty interval based on the adjusted variance of Eq. (10), is this correct?

In the Bayes linear framework used in this paper, the prediction takes the form of the expected value of f given the results of the collection of simulator runs, and the associated uncertainty is the variance of f given the simulator runs

Also, what is the link between Eq. (9)-(10) and Eq. (8)? Is the expected value E(B) in Eq. (9) given by the first term in Eq. (8) (the linear combination of "g" functions)? What about the variance? Please clarify.

We have added a specific example in the Bayes linear section of what E{B} and so on mean for the fast emulator. In the most general case, E(B) would be (sum E(beta)_i g_i(x) + E(u(x))), and Var(B) would be Var(sum E(beta)_i g_i(x) + E(u(x))) and hence include lots of correlations terms between the betas and u(x) which would have the be specified a priori. In our application, much of this is skipped by taken E(beta_i) as the regression estimates, Var(beta_i) as zero, E(u(x)) as zero, and Var(u(x)) as the residual variance, and leaving only Corr(u(x_1), u(x_2)) to be modelled (and even this is done using the data rather than specifying a prior).

- P. 19 L. 1 says that the Bayes linear approach is used for "the analysis of the link between the fast and slow emulators". Does this mean that Eq. (9)-(10) are also used to establish a link between the two emulators? If so, does this mean that the prediction of one emulator is adjusted based on the prediction of the other? This seems in contrast with Sec. 5.3, from which I understand that the link between the two emulators is established by linking their respective coefficients beta. Again this should be clarified.

Yes, there is a second Bayes linear adjustment used in the section. The link between the emulators is modelled as a link between the betas, and the Bayes linear adjustment is used to update this link based on observations. In the first paragraph of Section 5.3 this further use of Bayes linear has now been explicitly stated. The second paragraph in this section now also makes explicit that Bayes linear is being used to learn about the u'(x), the rho_i, and also the w(x) (that is, the local variation in the fast emulator, the link between fast and slow emulator coefficients, and the remaining local variation in the slow emulator).

- P.18 L. 4: "Conceptually, the expectation, variance, and correlation are a priori uncertainty judgements". What does this exactly mean? Which of the 3 parameters (expected value, variance and correlation length) are actually estimated from the residuals of the emulator predictions and which are assumed by a priori judgement?

Given the restructuring in the last revision, this sentence is not really meaningful anymore, so has been removed.

We've added some text (p17 line 10) to provide a general framework and make clear that the emulator structure may be designed using a mixture of judgement, exploration of the data and tuning. We've also removed the sentence referred to by the reviewer about the three specific parameters. The precise way the three parameters are chosen for our specific problem can't really be explained until some more ideas (Bayes-linear, fast and slow emulators) have been introduced, but we hope the added text will help the reader understand the range of possibilities.

- How is the "correlation length" estimated? The description in A2.2. is unclear: what does "tune" mean on line 25 P. 34? Manual tuning? How is it checked that the method "has been successful" L. 1 P.35?

We have expanded this section to explain the method more clearly. A subset of the observed data is used to find a correlation length that provides accurate estimates with as low variance as possible. The full data set is then used to check that the correlation length gives good predictions on the entire data set.

[3] From Figure 3 it seems that the difference between the fast and slow simulators is really small (at least for the chosen output variable). In the best case (top panel) the two simulators produce almost identical output, in the worst case (panel (b)) the outputs are still well related to each other (only few points in the bottom left part of Fig. 3.b would not align to a simple interpolating line). I guess this is the reason why the two emulators are found to be very close (P. 24 L. 4-6: "the link between mean functions of the two emulators is strong and consistent") and their difference "mostly a rescaling". I think the similarity of the fast/slow simulators should be emphasized more when commenting the emulator results, also to acknowledge that this case study application may not be the most challenging one to test the multi-level emulation approach (although the idea remains valid and very interesting in principle).

On P24 we have added "Given the similarity between the two simulators, it is not surprising that the multi-level emulation method works smoothly in this application; in applications with more fundamental structural differences between the simulators, it is likely that more careful modelling of the link would be required." To draw attention to this point.

MINOR

P. 4: maybe remove the line break between line 2 and 3 - the sentence "Finally, ..." should be part of the previous paragraph.

This has been corrected.

P. 4 L. 7-8: "The emulation method that is presented in this paper gives assessments of uncertainty that can be combined easily with other sources." I agree this is possible in principle but not actually demonstrated in this paper. Maybe good clarify the point.

This is a good point, we have added "(although actually performing such an assessment and combination is beyond the scope of this paper)" to this sentence.

P. 4 L. 10: "is expensive in both time and money." A bit vague. If I understand correctly, the point here is that sensitivity tests of a wide range of parameters require a lot of model runs and for a complex simulator such as VATD this would take a lot of computing time.

We agree this is vague and have added “to perform such an analysis requires very many simulator evaluations and hence very large computation time” to this sentence

P. 5 L. 15: "the emulators used..." remove "used"?

This has been corrected.

P. 5 L. 21-22: "Section 5 gives an overview of the statistical methods used in the analysis." Maybe specify this is about building and evaluating the emulators

We agree. This has been changed to “gives an overview of the statistical methods used to build and test the emulators.”

P. 10 L. 2: "all the alternative PSDs could be reconstructed to a reasonable approximation": unclear. What is "reasonable approximation"? Does it mean that the alternative PSDs are compatible with the range of observations by Dacre et al. (2013)?

We agree that this is unclear. The observed PSDs presented in Dacre et al. (2013) can all be reconstructed by varying the shape and scale parameters in a gamma distribution. The text has been updated to make this more clear.

P. 17 L. 14: "hosen" should be "chosen"?

This has been corrected.

P. 17 L. 15: "For the rest of this section, attention is restricted to scalar-valued for simplicity of notation." Ok but please clarify whether in your application a vector-valued emulator was used.

We agree this could be clearer and have added “For the application to NAME, f is vector-valued but this is handled by constructing separate scalar emulators for each f_i.”

P. 18 L. 20: "It" should not be capital letter

This has been corrected.