Response to Reviewers for Manuscript nhess-2016-288: "Multi-level emulation of a volcanic ash transport and dispersion model to quantify sensitivity to uncertain parameters"

Editor/reviewer comments are black. Responses to comments by the authors are red.

To address the reviews, some major restructuring and expansions are necessary. We observe in the two reviews that one reviewer asks us to provide more detail in some places, particularly about some aspects of the methodology, whereas the second review would prefer some of this discussion to be shortened where appropriate. To address both of these requests adequately, it should be possible to significantly tighten the main account, but rather than omitting some of the detail, instead move the some text to the Appendix.

Review 1

We thank the reviewer for taking the time to review our manuscript**.**

The subject manuscript describes an inversion of ash cloud transport in order to make a forecast. The manuscript is poorly written and organized, and is hard to read critically. It has one glaring weakness in that Bayesian linear regression is used on a high dimension problem formulated by the authors. Yet the emulator operates in a way to just constrain those parts of the model that are constrained by the data. This is a desirable property, but carries with it a liability. With this multi-level approach, there is the possibility (very likely) that there is a highly nonlinear dependence on the network function of the parameter values, in which case an exact Bayesian treatment is no longer applicable. The posteriors may be multi-model, and possibly non-convex. None of this is addressed, and it is also very difficult to follow exactly how the model operated on the May 14, 2010 ash clouds from Eyjafjallajokull volcano to provide the results in Table 4.

My recommendation is that the manuscript be completely reorganized and rewritten, being more explicit in some areas but also putting some explanations in Appendices or deferring to previously published work. It is very difficult to assess the efficacy and veracity of this study. Yet this is an important problem that deserves better treatment than is presented here. This could have been a shorter and much more informative paper than it is in its present state.

The linear regression component is only part of the analysis. When the regression is information, then the leading terms in the regression will identify some of the features of the simulator (this is why they are appearing in the regression). The features that are not captured by the regression are represented by the local variation in the Gaussian process residual. We use the validation techniques described in the paper to give us confidence that the emulator is valid and therefore that this approach is working. Validation over a large proportion of trial input values strongly suggests, by basic sampling arguments, that the conclusions are relevant over a large proportion of the input space. Of course, there may be specific areas where the emulator is performing poorly even though validation overall suggests the situation is acceptable, but in this study there were no extreme failures of validation (failing points were only a little outside the acceptable interval) and there was no obvious systematic pattern to these failures.

For the second part of this comment, the method of history matching hinted at in the conclusion provides a framework to address the highlighted issue. This is an iterative approach that progressively removes regions of the parameter space in which the emulator provides a poor match to observed data. In the remaining regions, the simulator is sampled again and new emulators are built within each region. This is a conservative approach that only rejects implausible regions. We are in the process of carrying out a full application of this method with NAME, but including it in this paper would make it unreasonably long.

A corollary of this history matching process is that after an application of history matching, the emulators must only be built in a subspace of the original parameter space, and there the form, as well as the coefficients, of the emulator can change. Therefore, as history matching progresses, the regression can include new terms that better model the behaviour of the simulator in these regions. Emulation is an iterative process, and this study performs the first steps.

It is, however, important to be clear what the scope of this paper is. Our analysis is focused only on understanding the behaviour of the NAME model itself, and the influences of its parameters, not its relationship with reality. The emulators we have built can be used to investigate this relationship, using the method of history matching hinted at, and we are performing this investigation, but including it in this paper would not be feasible. We have made this distinction clearer. Of course, as noted above, the insights into this relationship will be even greater after history matching has been performed.

Specific comments:
1) Use references with descriptions of how NAME works to shorten section 2.

Section 2 contains 3 paragraphs and provides the reader with information about NAME and the case study being considered. The authors believe that all this text is necessary.

2) page 5, line 24: particle density and particle size distribution are not known at the time of an eruption, so a forecast model must rely on a priori information.

We agree. Here we are highlighting the need for information about the volcano. Initially these ESPs not known or poorly known and a priori assumptions must be used until observational data is known. The text in Section 3 clarifies this.

3) page 13, line 17: only fine particles are considered, so why have particle size distribution as a variable, increasing the dimensionality, in this model (described in a later section).

In this context “fine” is not a constant value but a distribution between 0.1 and 100µm. While only the finer particles are considered, these are not all so fine that the size has no effect on the predictions. They need to be fine enough to reach the distal plume, but may still sediment significantly over the duration of the simulation. Here the particle size distribution is described using two parameters, shape and scale of a fitted gamma distribution which is a reduction from the use of 6 particle size bins.

4) page 14, line 20: 10,000 per hour, 1000 per hour are the rates of particle release, not the number of particles.

Agreed. The text on page 14, line 20 has been updated to reflect this.

5) page 16, lines 15-20: these are properties. So remove the word ’must’ and replace ’should predict’ with ’requires’.

Text has been updated to the following:

- It evaluates quickly
- It is expressive ...
- It predicts that ... are very close when ...

6) page 16: write out the function g, here equal to ’x’ to the ith power.

This is incorrect. i indexes the basis functions and, at this point, these functions could be anything. Later on (section 6.1) they are chosen to be the functions x_i, x_i x_j (i /= j) and x_i^2 with i,j ranging over the inputs. No powers higher than 2 are involved - there are a lot of functions because there are a lot of components to the input vector x, not sa lot of different powers. Near end of p16 we have changed the text to "Here, the $g_i(x)$ are chosen to be simple functions ..." to emphasise that the g_i are a choice.

7) page 16: define both expectation and correlation rather than just putting in E() and Corr() and expecting the reader to just know what this means.

Agreed. This has been clarified in the text.

8) page 17: same comment as above for Var() showing up suddenly in the text.

Agreed. This has been clarified in the text.

9) page 17: line 25: ’some of these problems are summarized’. Nice lead in, except that none of these problems are summarized later on. I suspect they got abandoned in an earlier draft.

The problems we had in mind are discussed on p18, lines 16-24 in the original manuscript.

10) ’expert judgements’ throughout are the a priori in your Bayesian analysis.

We agree (although the expert judgements can be hard to express in the form of probability distributions). We hope our agreement is clear in the passage from the last paragraph of 5.1 to the end of page 18 in the manuscript. However we have amended the text to make this clearer, referring to both "experts" and "prior" quantities in both sentences.

11) page 18: The paragraph/sentence starting with "Such an approach" is a non-sequitor with the paragraph and sentence immediately above. What approach?

We mean the approach that the previous paragraph describes - i.e. a full Bayes calculation. We are unsure why this isn't clear, but we have been more explicit with "A full Bayesian calculation of the type just described has been successful in many applications."

12) page 18, 1st paragraph: forecasting uses the posterior, which
has the prior in it. Change ’For calibration and forecasting’ to ’For calibration’

This is true in many Bayesian calculations, but things are more complicated with emulators. Bayes is being used to determine an approximation (i.e. the emulator) to the function f(x) representing the dependence of NAME's output on certain parameters. Here the NAME simulations are the "observations" which lead to the a posteriori model. At that point we have an approximation to NAME, but values of the parameters still need choosing if we are to make a prediction for concentration levels. This could involve a separate Bayesian calculation using e.g. satellite observations of the ash cloud. This is what the text very briefly refers to. In fact the paper only uses the emulator to better understand NAME's sensitivities and so the text here (last sentence of the paragraph) could be removed. However we think it helps give some wider context to what is being presented.

13) page18: The reason that doing high dimension problems with a linear Bayesian analysis is very difficult is that linear regression is not well suited to high dimension problems. You don’t have enough flexibility with fixed basis functions. This could have been alleviated in this paper by just considering two parameters for distal clouds; height and mass flux at that height.

We mostly agree, but considering just two parameters would be a different paper with a different aim. Our aim here is to explore the uncertainties and sensitivities of NAME arising from a wide range of uncertain parameters. It is certainly possibly that we might not have had enough flexibility in the basis functions (which would put more importance on the residual functions u and u’), but this can be tested after the event by checking the emulator output against some NAME simulations which have not been used in the emulator.

14) page 20, first paragraph: so, a prior was not used here?

That’s correct. To avoid any doubt we have changed "standard (non-Bayesian) linear regression" to “... linear regression (without a prior)”.

15) page 22: give references for R squared.

This is a very standard statistic as such it is not normally given a reference. Text has been updated to "The coefficient of determination, R^2, which represents ...".

16) page 24, line 24: change "..., whereas of course" to "..., even though".

This change has been made in the text.

17) page 24, bottom: so, the number of parameters could not be reduced below 4? What if you did not do this? Would you just end up with height and mass flux? I suspect so.

If we removed the requirement to have at least four parameters in each model, the results would be as follows. In many of the regions the process would still select four parameters because the regressions were poor with fewer than four parameters. In the remaining regions, most would be reduced to 2 parameters and indeed these would be height and mass. In a couple of instances, a different parameter to mass would be chosen (height would be in every model), or three parameters would be chosen. However, most of the 2- and 3-parameter emulators would not pass validation.

18) page 34, Cov equations at top: there are 2 parameter sets here. Need either references or a proof of the second equation.

This is a consequence of Cov(\Sum_i X_i, \Sum_j Y_j) = \Sum_{i,j} Cov(X_i, Y_j). However we agree that the text needs to be modified to make it clearer to the reader that u' is not random here and that we are considering only x_j in \chi_F (and so u' is known).

Review 2

The manuscript presents a complex and potentially very interesting application of emulation modelling to quantify the relative impact of several uncertain parameters on the predictions of a volcanic ash transport and dispersion model (the NAME model). The main novelty of this study is that it presents one of the first applications of a formal sensitivity analysis (i.e. beyond one-at-the-time approaches) to a volcanic ash transport and dispersion model. Moreover, some of the techniques and lessons learnt in this study (for example on how to link fast and slow simulators) might be of interest to a broader community who deals with uncertainty in computationally expensive transport and dispersion models, not only volcanic ashes.

However, the manuscript in its present form is quite unclear and the structure unbalanced, which makes it difficult to fully evaluate and appreciate the findings of the study.

We thank the reviewer for taking time to review our manuscript and give such constructive feedback. We agree that the paper could be reorganised to make it clearer to the reader the key messages of the study.

Below my main issues and some suggestions for revision.

[1] The choice of contents in the methodology sections is confusing. Too much space is given to topics that are generic and covered in many other papers or even textbooks, which divert attention from and leave too little space to specific information about the NAME application. For example, on P. 14 L. 16-29 almost the same space is devoted to describing the difference between fast and slow simulators, which is a very important setting of the model under study (see also comment [3] below), as to describing the well-established and totally generic Latin Hypercube sampling technique.

We agree and the long section outlining the Latin Hypercube sampling technique has been reduced.

The description of emulators and Bayes linear methods covers 4 pages (P. 16-19) although much of the content is very generic, easily accessible elsewhere, and - most importantly - not strictly related to the application here. In fact, from P. 20 L. 10-12 ("a successful method has been to use a standard (non-Bayesian) least-squares regression") I understand that the non-linear/linear Bayesian approaches discussed on previous pages are not actually used here because linear least-squares are sufficient.

The Bayes linear calculations are used later (see the next comment), therefore the Bayes linear section needs to be included but it will be revised to highlight the points needed to follow the narrative of the paper.

As a reader I was confused by these long descriptions and diverted from focusing on the specific features of this application, for example the link between emulators of the fast and slow simulators. Another example is Section 5.4: this is all generic and standard methods that do not need to be discussed in an application-oriented paper, especially if not used then in the Results section (see for example the comment on "examining the residuals"...). In summary, I would recommend to deeply revise these sections, to make them shorter and include only the information relevant to the specific application and thus needed to understand the results.

We agree that the balance should be improved. As noted by the reviewer, the two main focuses of the paper are, or should be, the application to NAME and the use of multiscale emulation. The other parts of the methodology sections can be significantly shortened to improve clarity. To further address one point in this comment, “I understand that the non-linear/linear Bayesian approaches discussed on previous pages are not actually used here because linear least-squares are sufficient”, while it is true that we use the least-squares fit to fix most of the parameters of the emulator for the fast simulator, there is still a Bayes Linear adjustment performed to adjust the emulator for the slow simulator. Therefore, some discussion of Bayes linear is necessary, and has been reduced (and we specifically clarify exactly where the Bayes linear adjustments are happening and where they are not).

However, since the first reviewer would like some expanded detail in some of these areas this material has not been removed but moved to the Appendix.

[2] The results section seems to focus a lot on the validation of the emulators - i.e. how good they are in representing the fast and slow simulators - and their similarity, rather than the ultimate goal of the analysis, which is to use the emulators "as a research tool to better understand the simulator, the role of the parameters, the interactions between them..." (P. 4 L. 17). For example, while Fig. 5 and 6 report validation results for the emulators, there are no Figures reporting sensitivity results, parameter mapping or interactions. The only results related to sensitivity analysis are in Table 4, which however does not even use the word "sensitivity"! This is odd especially considering the title of the manuscript.

We agree that there is a large focus on the validation of the emulators. However, it is important to bear in mind that the emulation procedure is based on a lot of choices (basis functions, correlation function, correlation distance, and so on) and that it is the validation step that gives confidence that the choice made were good ones. Therefore we need to make sure we include enough information that readers can have confidence that the emulators actually work. To reduce the amount of space dedicated to this in the paper Fig 6 has been removed and some details of the validation process has been placed in the appendix.

We have added the specific details of the expectations and variances of the adjusted coefficients in the model, the ranges of predictions that can be generated from the parameter ranges. Unfortunately we have not been able to include plots of the emulator predictions at this time.

[3] The difference between fast and slow simulators should be better clarified. In particular, on P. 14, L. 16-24: What does the "increase in particle-sampling noise" mean in simpler terms? How does it impact the simulation results? Not so much - at least for predicting average column loadings - according to what the authors say later on P. 15 (L. 7-12).

Particle sampling noise is due to the stochastic motion of the particles within NAME and the consequent randomness in the number of particles in a given region, and hence in the total particle mass in a given region. The fractional noise is proportional to 1/sqrt(N) where N is the number of particles. Here the impact is small as we average over large regions. Text has been added to the manuscript to clarify this.

On the other hand, how fast is the fast simulator?
These aspects need to be clarified because they are at the basis of the motivation of the study: if the fast simulator is fast enough, and its predictions of the output of interest (average column loading) are reasonably close to those of the original (slow) simulator, then why building an emulator? One could directly apply a Global Sensitivity Analysis technique (for instance, the Morris method, or Regional Sensitivity Analysis, which are both reasonably "low-cost") to the fast simulator. I am not saying that this is necessarily the case (maybe the fast simulator is not so fast, or there is some other reason I am missing to avoid using it) but more details should be provided to clarify this crucial point.

The fast simulator takes between 10 and 20 minutes to run. So much faster than the “slow” simulator but not quick enough to apply the techniques you suggest. Text has been added to clarify this.

[4] Choices underpinning the analysis and their impact on sensitivity results needs further discussion. P. 6, L. 20 onwards: there are three set of choices that are made here: - the choice of the uncertain parameters to be included in the sensitivity analysis (out of a larger set of parameters appearing in NAME, which are set to default values - two of them are further described in Sec. 3.2.7 and 3.2.8, together with the reasons for not varying them - what about the others?) - the choice of plausible ranges for the

uncertain parameters - the choice of a particular event, and hence a particular set of forcing data, for the model simulation (and the choice of ignoring the uncertainty in those data) I presume that these choices may have a very strong impact on the results and thus on the generality/transferability of the findings. Assessing such impact might be beyond the scope of this manuscript, but the point should be at least mentioned and discussed.

In this study the input meteorology was not varied because the large dimensionality of the met input makes it unsuitable for this type of approach. Also at the time an appropriate set of ensemble meteorology was not available. We accept that this could have a large impact on the ash column loadings, although in the case study presented the meteorological situation is relatively settled. However this does not reduce the value of understanding the uncertainty due to other causes, while recognising that this is not the total uncertainty. Text will be added to the conclusions to clarify this beyond the scope of the paper. Expert elicitation identified the parameters leading to uncertainty and where a parameter is set to the default value it is highlighted.

The parameters chosen to be included in the sensitivity study, along with their plausible values, were determined through an expert elicitation exercise. Full details of this procedure are beyond the scope of this paper but they involved consideration of estimates for these parameters found in the literature, choices made by various experts who run this and similar models, and from the personal experience of the co-authors from the Met Office who have significant experience with NAME. A brief discussion of this can be added. If the reviewer considers this aspect particularly important, then we would be happy to provide an expanded discussion in the appendix. Section 3 presents these parameters and where a parameter is set to the default value this is highlighted. From the perspective of understanding the sensitivities, the exact ranges are not crucial. The ranges are more important if we try to infer uncertainty directly from the ranges, but the intention is not to do that but to infer more suitable ranges from history matching.

We accept that we have only presented one case study and that this has limitations although we feel we have not brushed over this. The conclusions already contain a sentence “These conclusions should be tested in other situations to assess how widely they hold” (Page 30 L16). The same paragraph also starts with “For this case”.

OTHER SPECIFIC POINTS:
P. 3 L. 28: "Finally, the analysis cannot provide ...". A bit vague, please clarify what is an "overall assessment of uncertainty"?

The uncertainty from the use of a computer simulator comes in very many forms, for example parameter uncertainty, measurement uncertainty, uncertainty about the effects of assumptions within the model, uncertainty about the impact of missing processes, and so on. The primary uncertainty being considered in this paper is the uncertainty associated with the value of the simulator at a set of parameters at which it has not been run. The simple analysis being discussed in the introduction does not lend itself to a numerical representation of this uncertainty. In contrast, the emulation method contains exactly such a representation, which could then be combined with all other assessments of uncertainty to provide an overall assessment of uncertainty.

P. 8 L. 3-4: Difference between "full depth" and "thin layer" is not clear to me. Also, very unclear how the 1700 + 1700 runs here mentioned are connected with the 1500 + 200 runs mentioned on P. 14 (the same experiment of P. 14 is repeated twice, once per each source type?). Maybe the confusion could be avoided by simply not giving all these details on the thin layer source case, since its results are not shown (as commented on P. 23 L. 15-16)?

When initially designing the study two sets of runs were performed to test the impact of the vertical distribution of ash within the ash column. The majority of the references to this have been removed and replaced with a "the same analysis was performed with the thin layer source, with no significant differences in conclusions (although the emulators in this case performed slightly worse consistently)".

P. 10 L. 18: " are varied by the same proportion". Unclear.

We agree this is unclear. The text has been revised to clarify this.

P. 12 L. 23: "between 0 and 2": or between 0.5 and 2 (assuming this section refers to x_{17} in Table 1).

The scaling factor does vary between 0.5 and 2. The text has been updated to reflect this.

P. 14, L. 10-11: "the average ash column loading predicted ..." not completely clear. Are column loadings per each region averaged over the simulation period, thus defining 75 "outputs" (and 75 emulators), or are predictions for each hour analysed separately (thus defining 75xT "outputs", where T is the number of hours in the simulation)? Please clarify, maybe also inserting an equation here.

There are 75 emulators in total. There are up to four regions each hour for the 24 hours simulated. The regions are chosen using the satellite retrieved ash column loadings so the regions change from hour to hour. All regions are shown in Table 3. The text has been revised to make this clearer. Note that the NAME column loadings are averaged over the hour (not snapshots).

P. 14, L. 14: Again unclear: "regions used for the first hour are marked..." So, the definition of regions changes from one hour to another? And also their number then? And so how does it connect to my previous question?

Please see comment above.

P. 15, L. 7-12 and Figure 3. Again on the difference between fast and slow simulators. I understand that the main conclusion here is that simulations from fast and slow simulators are similar, however this paragraph and the Figure are rather unclear. "agreement between the two simulators" means that they provide similar predictions of average column loading? What does "to be related but not in agreement" mean? The "correlations" of 0.99 and 0.7 are the correlation between simulated column loadings (in different regions? at different time in the simulation period?). Please be more specific. Figure 3: units of measurements on the axes are missing!!! (I guess they are "Log ash" as in Fig. 5?).

The correlation range is the range over all 75 regions that are emulated over the whole simulation. Each correlation refers to a single region and time, and that the correlation is over the 200 different choices for the parameters. There are 200 points in Fig 3a and 3b. The text has been revised to make this clearer and the "related but not in agreement" statement has been rephrased.

The intention was to say that for some of the regions there is an almost exact match between fast and slow simulators, whereas in others there are noticeable differences between the predictions of the two models but there is still a clear relationship (this can be of two forms: in some regions, if one parameter set has higher output than another in the fast simulator, it will have higher output in the slow simulator, but the values will be different; in other regions, this is usually the case but sometimes there will be lower output (e.g. the second plot in Fig 3). The highlighted sentence is awkward and has been clarified.

The axes labels in Fig 3 have been updated.

P. 15, L. 7: "the goal" ... This is a bit misleading: the ultimate reason for building the emulator is not making inference, but rather understanding the role of parameters (see discussion on P. 4 L. 14-17). Maybe good to remind it here.

Yes, we agree this is a bit misleading. This is referring to the goal of the statistical technique not the overall study. The text has been revised to clarify this.

P. 28, L. 9-10: "for some parameters... it is not plausible": so you included in the analysis some parameter combinations that are implausible? This sounds contradictory (who would be interested in sensitivity to parameter variations that are not plausible?). Please clarify

We agree that it is not sensible to investigate parameter sets that are not plausible. This sentence was included to reiterate that the maximum turbulence values used in this study are representing plausible extreme values of turbulence (P10 L16-17). It is not expected that these values would be present everywhere in the atmosphere but the current model parameterisation uses a constant value of turbulence in the free troposphere. To model spatially/temporally varying a new parameterisation would need to be developed. Text has been added to clarify this point.

P. 29, L. 18-23: This is a generic comment that was already made in the Introduction (P. 3) and is not part of the results. I would remove it.

We agree. This section of text has been removed.

Table 1: - term "default" on first row is a bit misleading - does it refer to the "data

from Keflavik radar" (as explained on P. 7, L. 19)? If so, clarify in the Table - term "default" on third row is a bit misleading - does it refer to the MER value from Eq. (1) (as explained on P. 8, L. 20)? If so, clarify in the Table - connect better to the text, for example the mathematical symbols used in the text could be included in the description of "Parameter name" (for example R_a on row 17...) - rows 7-10: "varied in proportion..." is unclear (here and in the text) - row 18: so the default value coincides with the maximum value? This looks strange.

Default in row 1 will be replaced with “Arason et al.”.
Default in row 3 will be replaced with “Mastin et al. “
As suggested the mathematical symbols for the parameters will be included in the “Parameter name” column.
For clarity text “varied in proportion with” will be removed from the body of the table.
The default value being set as the maximum value is what was decided upon during the expert elicitation exercise.

Figure 3: Caption mention "(a)" and "(b)" but letters are not reported in the panels.

This inconsistency has been corrected in the revised manuscript.