Dear Editor,

We thank you for having our paper reviewed. We agree with the reviewers' comments and have addressed each of them in the revised manuscript. The changes are highlighted in red in the manuscript. Below is our response to each specific reviewer comment.

Interactive comment on "Spatial evaluation of volcanic ash forecasts using satellite observations" by N. J. Harvey and H. F. Dacre

M. Fromm (Referee)

mike.fromm@nrl.navy.mil

Received and published: 8 October 2015

This paper introduces a technique for assessing skill in volcanic ash (VA) forecasts or simulations that has an advantage (as the authors argue) over traditional point-focused methods. That advantage is realized in the case that the evaluator's goal is to quantify similarity in the horizontal footprint of an observed VA plume. My understanding of the problem is that it is important to quantify how well a simulated VA plume matches the size, shape, and location of a VA-plume retrieval by an imaging satellite (e.g. SEVIRI or MODIS). In a range of scores reflecting agreement between observation and simulation, from 0% (no agreement) to 100% (perfect agreement), there is said to be skill in the simulation at 50%. Hence this technique offers a continuous skill-score range and a skill/no-skill threshold for the purpose of comparing different simulation assumptions or meteorological drivers within one model, different volcanic source terms, or model- model differences. As such, this topic and the authors' manuscript are appropriate for ACP and potentially beneficial to the community striving to improve VA forecasts and warnings.

The manuscript is well written and organized. It was easy to follow and understand. The results consist of a single demonstration of the proposed VA evaluation technique, which is necessary but perhaps not sufficient convince the reader of the utility and value of their spatial evaluation technique. Hence I would recommend this paper be published after consideration of adding at least one more example of a real VA plume simulation, and satisfactory response to the minor and technical comments listed below.

To address this we have added 3 more examples of real VA situations based on the same NAME simulation.

P24728, L18. Consider amending "The presence of. . ." to "The presence or threat of. . ."

We have amended this as per the suggestion.

P24733, L22. Please consider informing the reader of the initiation time of the NAME simulation shown and discussed here. IT may not be critical to the paper, but I was left wondering how many hours/days post initiation these results were.

The simulation start date is 1 May 2010 so the results shown here are on day 14 of the simulation. This has been clarified in the text on P24731 and P24733, L22.

P24734, L14. "satellite retrieved" should be hyphenated I think.

We agree. This has been updated here and a number of other places in the text.

P24734, L18-24. Here the authors introduce the pixel-matching concept. An issue came to mind while reading this that may represent a flaw or at least some incompleteness in the pixel-matching construct. It seems to me that the number of pixels with a simulated VA amount exceeding the ash/no-ash threshold is directly tied to the source VA concentration. This has nothing to do per se with the VATD model itself. E.g. a relatively small initial dose of VA will lead to a relatively small plume at all forecast times. I would expect that this would or could impact the number of matching pixels but this plume size is unrelated to the model itself that is being assessed for skill. The authors discuss how an artificial cut off in the satellite-ash retrieval affects the pixel-matching process but not the point raised here. Hence it would seem that there must be a discussion of this if indeed the authors agree that there is merit to it.

We agree that for eruptions that emit small amounts of ash that this technique may not be suitable as the ash levels may be below the detection limit of the satellite. In this case the VATD model would not be penalised as pixel matching would match the number of simulated grid boxes to the number of observed grid boxes (which could be zero).

P24735, L9. “. . .gridbox are. . .” should be “. . .gridbox is. . .” I think.

We agree. This has been updated in the text.

P24737, L26. “assess” should be “assesses”

We agree. This has been updated in the text.

P24739, L2-3. This sentence is unclear because it is a comparison without invoking two points of comparison. It states that the "objectively...results" are "more similar" to "subjective...inspection" but doesn’t say with respect to what.

The text has been clarified here.

P24739, L17. Hyphenate “satellite-retrieved” Figure 4 legend. More space is needed between the dots to show the differing line types.

The hyphenation has been updated and the lines in this figure are now different colours to alleviated the problem of the dots obscuring the differing line types.

Interactive comment on “Spatial evaluation of volcanic ash forecasts using satellite observations” by N. J. Harvey and H. F. Dacre

Anonymous Referee #3

Received and published: 24 November 2015

General Comments

In this new paper Harvey and Dacre discuss the evaluation of volcanic ash forecasts with the NAME Lagrangian particle dispersion model using SEVIRI satellite observations. The fractions skill score (FSS) is used as performance metric. It evaluates the spatial extent of the ash distributions from the model and satellite data. Choosing different neighbourhood sizes, the FSS approach allows to determine the spatial scales on which the model has forecast skills. This is demonstrated for a case study for the 2010 Eyjafjallajökull, Iceland eruption. I found this paper interesting to read. The FSS method is interesting to other researcher trying to validate their transport model simulations. The topic fits in the scope of ACP. The paper is very concise and mostly well written. However, I have two general comments and a number of specific comments. I would recommend the paper for publications once these comments are carefully addressed.

1) In some places this paper reads as if it introduces the FSS as a completely new performance metric. However, it was already established in earlier work (Roberts and Lean, 2008), in the context of validation of precipitation forecasts. It might be the case that this is the first paper that applies the FSS method for volcanic ash forecasts? At least, I did not found any other papers using it for that purpose. I added specific comments regarding this issue below. As this issue relates to the main aim of the paper, please clarify this.

See responses below

2) The paper claims that the FSS is more suitable than "traditional" point-by-point performance metrics such as the critical success index (CSI) or Pearson's linear correlation coefficient (PCC), because it permits to assess forecast skills on different spatial scales. However, this is a rather general conclusion that is drawn based on just one day of data from one case study. Based on the data and results presented for this one day, I was not fully convinced that the FSS really is a more suitable performance metric than the CSI or PCC. Analyzing at least 2-3 different days of the Eyjafjallajökull with a comparative approach might be helpful in order to support the conclusions. It would be interesting to see how the NAME forecast performance changes during the course of the simulation.

To address this we have added 3 more examples of real VA situations based on the same NAME simulation. The temporal evolution of FSS during the Eyjafjallajökull is the focus of a paper submitted to Journal of Geophysical Research which is currently under review.

Specific Comments

p24728, l4-6: Here you might clarify that the FSS is an already existing metric, which you apply (possibly for the first time?) for the evaluation of volcanic ash forecasts.

Text has been added to clarify this point.

p24728, l8: I thought the "success index" is more commonly referred to as "critical success index" or "threat score"?

Yes the success index is often referred to as CSI or threat score. However in volcanic ash literature it is often called success index (e.g. Stunder, 2007; Webley at al., 2009).

p24728, l17-18: A reference might be added, e.g., [Casadevall, 1994; Miller and Casadevall, 2000; Prata, 2009].

We have added references to Casadevall, 1994 and Miller and Casadevall, 2000 here.

p24729, l2-7: In fairness, you might also point out advantages such as high temporal and spatial resolution of the ground-based or airborne in-situ measurements here?

We have pointed out the high temporal resolution of the ground-based and airbourne measurements here although these observations do not have high spatial coverage..

p24729, l9-10: High temporal resolutions is only available for geostationary satellite instruments (e.g., SEVIRI), but not for satellite instruments in low Earth orbits (e.g.,AIRS, IASI, OMI, ...), which are also frequently used to study volcanic events.

Geostationary has been added to the text to clarify which instruments we are referring to.

p24730, l26: Add a reference for SEVIRI?

A reference to SEVIRI has not been added. This is in agreement with many other papers which use data from SEVIRI (e.g. Francis et al. 2010 and Millington et al., 2012).

p24730, l25-27: Here you might explain why you picked just one day for your case study? Why did you pick 14 May 2010, specifically?

As per your general comment above we have now extended the number of case studies presented. Originally this case was chosen due to the large number of supplementary in-situ measurements available.

p24731, l2-3: Is there a reference for this first application/case study of the NAME model?

Reference to this is in the introduction of Jones, 2007.

p24731, l5-7: Is there a reference for the UK Met Office global NWP analysis?

A reference to this is not usually given (e.g. Witham et al. 2012, Dacre et al. 2011)

p24731, l7-9: Are there references for the physical schemes (turbulent diffusion, sedimentation, dry deposition, etc.) used in NAME?

A reference for this has been added: Witham et al. (2012) The current volcanic ash modelling setup at the London VAAC found at http://www.metoffice.gov.uk/media/pdf/p/7/London_VAAC_Current_Modelling_SetUp_v01-1_05042012.pdf

p24731, l13-16: Is there a temporal development in the eruption source parameters considered in the NAME simulations?

Yes, there is temporal variation in the plume rise height as per the observations in Arason et al. 2011. Text has been added to highlight this.

p24732, l27-p24733, l4: The SEVIRI data are averaged for a 5h time period. What are the spatial scales correlated with this time period? Does this have any influence on the scale analysis performed with the FSS?

We have calculated the correlation between subsequent satellite images for varying time periods and the results are consistent with the RMSE analysis. The analysis has been performed without the 5 hour averaging and for the cases presented here the FSS results are very similar.

p24733, l18-19: This might be another place to add some information why you picked this specific day for the analysis. Perhaps it should also be mentioned how many days after the eruptions it is, as forecasts skills likely vary during the course of the simulation?

We *have extended the number of case studies in response to major comment 2. Text has been added noting it is day 14 of the NAME simulation and day 31 of the eruption. The time evolution of the FSS for this case is the focus of another paper which is currently under review.*

p24734, l2-4: Is the large dispersion of the volcanic ash cloud seen in the NAME simulations considered to be realistic? Is there any observational evidence for this?

The simulations presented here are consistent with those performed by Grant et al. (2012), Devenish et al. (2012) and Heinold et al. (2012). Validating the dispersion over such a large domain is difficult – hence the motivation for utilising the new satellite retrievals in this paper. Although NAME is state-of-the-art it is reliant on the input meteorology, small errors in the wind speeds in this input data can *hugely impact the long range dispersion. Also, the input meteorology affects the wet deposition through precipitation. It is also possible that the missing aggregation processes cannot be simply accounted for through the DFAF. Note that the scale on Figure 1 starts at 10^{-2} micrograms per $m^{2.}$ Assuming an ash layer depth of 1km, this is a very small concentration value.*

p24734, l14-16: Can you be more specific and provide an error range of the satellite ash column data? Perhaps replace "Therefore the values can be considered..." by "We considered the values..."?

The text has been updated as per the suggestions. At present we do not have robust error estimates for the satellite data.

p24734, l28-25: Instead of using the term "pixel matching", it might be more clear to state that you are making the forecast "bias-free"? By applying a time-varying threshold for the model it is ensured that the model never under- or over forecasts the observations. On the one hand this improves the performance metrics. On the other hand it may hide problems with the model (as relationship between the model and satellite ash column absolute values cannot be established anymore).

Clarification of the removal of bias has been added to the text but the term pixel matching has been kept.

p24734, l25-26 (and Fig. 2b): What do we learn from the time variations of the domain fraction?

When FSS is used in the validation of precipitation forecasts a set threshold is used. Here we match the number of NAME pixels to satellite pixels to remove bias so it is not possible to use a fixed threshold. Figure 2b shows the time variation of this threshold and the corresponding ash column loading. The figure is included to show that although the threshold is time varying it only spans a small range (85.4-96.6%). This range is cited in the text.

p24735, l3-5: I do not understand the relevance of the DFAF for your study. Is it a model parameter for the NAME simulations? Is there any large uncertainty related to it regarding the simulations?

DFAF is a measure of how much of the ash that is emitted from the volcano undergoes long range transport. It is applied to the output of the simulation and represents the fraction of ash that remains in the atmosphere after sedimentation of large particles (>100μm) and aggregation of smaller particles has occurred. The value of DFAF is uncertain but work done by Dacre et al. (2011, 2013) and Grant et al. (2012) suggests it lies within 2 - 7% during this phase of the eruption. Although previous studies (e.g . Rose et al. (2000)) of different volcanoes found slightly lower values (0.7-2.6%) This is why a range of DFAFs are used in Figure 2.

p24735, l8-11: I got confused regarding the neighbourhood sizes. From Sections 2 and 3, I thought you are analysing NAME data and integrated SEVIRI data on a 0.375◦ x 0.5625◦ (40 km x 40 km = 1600 kmˆ2) horizontal grid. This is much larger than neighbourhood sizes of 40-1160 kmˆ2 referred to here?

We think that the notation has confused the reviewer. In this paper 40 km^2 refers to a 40 km x 40 km region (i.e. the grid scale). To remove confusion km^2 has been replaced by $(km)^2$ in the text.

p24735, l13-p24736, l2: Even though references to the literature are already provided, it would be helpful if you could provide more details on how the FSS is actually calculated. I understood that the analysis is starting from the gridded model and satellite data. Then you are selecting (squared) neighbourhoods of N = n x n grid boxes for the analysis. In each neighbourhood j, the fractions O_j and M_j refer to the numbers of grid boxes where the observation and model thresholds are exceeded. Is this correct? Are the neighbourhoods distinct from each other or are they shifting windows?

Yes, your understanding is correct. The neighbourhoods are shifting windows.

p24736, l6-10: Again, I do not understand this minimum neighbourhood size of 40 kmˆ2. Perhaps you could also mention the FBS and FBS_ref values for this example?

See comment above. We do not think quoting FBS and FBS_ref is informative to the reader in this instance.

p24736, l19-24: Why did you apply this specific transformation? Why did you couple the stretch/squash factors in longitude and latitude, rather than using two distinct parameters for both directions? If the same stretching factor s is applied in longitude and latitude, the actual Cartesian distances (dx and dy) would be scaled differently, depending on latitude. Is this desired? How do we know which values for s would be realistic? Is the range from 0.5 to 2.0 tested here reasonable? Wouldn't it be more reasonable to perform this kind of test with different NAME simulations using different ESPs (such as different plume height, for instance)?

The transformation is idealised and has been chosen for its ease of application over performing simulations with different ESPs. The simulation presented is 14 days long and thus there are changes in both the plume height and input meteorology which makes it very difficult to design an experimental setup which would provide ash distributions that are very different from the one presented. Many other more complex transformations could have been applied however, we believe this one enables us to illustrate the spatial evaluation method.

The value of s is not directly being used to represent a physical process. Any value of s could be used, the ones chosen produce plumes that are still within our domain of interest.

p24737, l8-16: I was wondering if the neighbourhood sizes with skillful model forecasts found here are directly related to the stretch factor s? For instance, if you use a stretch factor s=2.0, can we expect that the neighbourhood size also grows by a factor of 2?

No, this is not the case (see Fig 4). If this was true then using a factor of 0.5 would give a neighbourhood size reduced by a factor 2.

p24737, l23: "critical success index (CSI)" might be more common than "success index(SI)"? A reference would be (Schaefer, Weather and Forecasting, 1990).

The reference to Schaefer (1990) has been added to the text.

p24738, l9-10: What are the actual PCC ranges and CSI values considered to be skillful in the papers of Kristiansen et al. and Webley at al.?

The SI values found in Webley et al. range from 0.17-0.60. Stunder et al. suggest an acceptable forecast has a SI greater than 0.25. Kristansen et al. state that PCC values of 0.36-0.48 are significant. These values have been added to the text to aid the interpretation of the results presented in this study.

p24738, l11-18: It is stated that "by visual inspection the stretch factor 0.5 ash cloud appears to more closely match the satellite retrieved ash than the stretch factor 2 ash cloud". However, this does not become evident to me from Fig. 3. Could you please explain in more detail how you made this judgment? Based on this single example, I am not convinced that the FSS works much better than the traditional point-by-point metrics, I am afraid. Additional examples could be helpful.

Visual inspection is very subjective. The authors believe that the stretch factor 0.5 ash cloud appears to more closely match the satellite retrieved ash than the stretch factor 2 ash cloud as it seems to overlap the satellite location more. As suggested further examples have been added.

p24738, l22-25: However, the CSI or PCC analysis could also be performed on different grid box sizes and therefore could also provide information on different spatial scales. Is there any reason why this would not be appropriate and the FSS should be used instead?

FSS is routinely used in the verification of precipitation forecasts. It provides an assessment of forecast skill over a range of neighbourhood sizes and displacement errors. CSI and PCC could be computed for aggregated grid boxes. However, to do this you would need to consider fewer "boxes". Also, this would still suffer from the same double penalty problem and not provide any information about displacement errors

p24739, l5-14: However, most of these analyses could also be performed with other skill scores?

Yes, we agree that this comparison could be done with any other skill score (and indeed is!) but this score specifically focusses on the spatial scale over which the simulation/forecast could be "trusted" and can be used to provide further information that traditional point scores do not.

p24742, l18-19: A web link to access the Oxford Economics report would be nice.

This has been added to the reference.

Fig. 1: The color bar range extends from 10ˆ-2 to 10ˆ7 ug/mˆ2. However SEVIRI cannot measure anything below 0.2...1 g/mˆ2 (Section 3). Is it useful to show model data for 6-7 orders of magnitude for which the satellite instrument cannot provide information? This may give a miss-leading impression that the NAME simulation is much too dispersive? A more colorful color-scale could help to infer actual values from this plot.

The color bar range has been chosen to show the extent of ash in the NAME simulations. The same color bar has been used in the satellite plot to be consistent. Panel c shows the pixel matched ash. As suggested a more colourful scale has been used.

Fig. 1: What is the reason for the increased amount of ash at 30◦ W, 55◦ N in the NAME simulations?

The ash at 30◦ W, 55◦ N is a consequence of "older" ash recirculating around a high pressure system located to the south of Iceland in the NAME simulation. This is the subject of another paper which is currently under review.

Fig. 3: A different color (e.g. red) for the satellite data contour line would be nice.

This has been noted and the plots have been updated.

Technical Corrections

p24729, l14-16: "The large spatial coverage ... over a large spatial scale." sounds a bit redundant.

We agree this sentence is not very informative. It has been revised.

p24729, l20: "sqaure" -> "square"

This has been updated.

Fig. 4: "Stretch factor: 0.7" -> "Stretch factor: 0.5" (in the plot key)

This has been updated.

Fig. 4: Example lines in plot key are too short.

The lines in this figure are now different colours to alleviate the problem of the dots obscuring the differing line types.

Figs. 2-4: There is no need to explain/repeat the line types and symbols in the caption, if a key is already provided in the plot.

This has been noted although this practice is not uncommon.

Interactive comment on "Spatial evaluation of volcanic ash forecasts using satellite observations" by N. J. Harvey and H. F. Dacre

T. Chai

It is a very nice short paper. Here are two short comments.

Page 24733, lines 2-4, It states that "The choice of a 5 h aveaging time was based on the results of some simple data denial experiments". However, it is not clear to me how 5 h averaging time was chosen based on Figure A1. That does not correspond to the minimum RMSE for any of the days.

The x-axis of Figure A1 has +/- before each number. The minimum is at +/- 2 hours for many of the days considered. This gives a 5 hour average as there is the current hour plus the two hours before and the two hours after.

Page 24740, lines 6-7, The statement, "This is because as the averaging window increases the amount a removed pixel contributes to the RMSE reduces", requires a little elaboration.

We agree that this statement could be more precise. It has been replaced with "RMSE penalises variance as it gives errors with larger absolute magnitudes more weight than errors with small absolute values. It is thus sensitive to outliers, which are reduced by the time averaging method. "