Atmos. Chem. Phys. Discuss., 15, C9695–C9700, 2015
www.atmos-chem-phys-discuss.net/15/C9695/2015/

Atmospheric Chemistry and Physics
Discussions
Open Access

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.

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.

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.

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

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

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?

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.

p24730, l26: Add a reference for SEVIRI?

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?

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

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

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

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

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?

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?

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?

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..."?

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 overforecasts 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).

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

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?

p24735, l8-11: I got confused regarding the neighbourhood sizes. From Sections 2 and 3, I thought you are analyzing 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?

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?

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?

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)?

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?

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

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.?

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.

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?

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

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

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.

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

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

Technical Corrections

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

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

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

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

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.

Interactive comment on Atmos. Chem. Phys. Discuss., 15, 24727, 2015.