Procedia Computer Science
Volume 29, 2014, Pages 1227–1235
ICCS 2014. 14th International Conference on Computational Science

Fast Construction of Surrogates for UQ Central to DDDAS – Application to Volcanic Ash Transport *

E. R. Stefanescu[1], A.K. Patra[1], M. Bursik[2], M. Jones[3], R. Madankan[1], E. B. Pitman[4], S. Pouget[3], T. Singh[1], P. Singla[1], P. Webley[5], and D. Morton[5]

[1] Department of Mechanical & Aerospace Engineering,
University at Buffalo, Buffalo, NY 14260, USA
{ers32,abani,tsingh,psingla}@buffalo.edu
[2] Department of Geology, University at Buffalo, University at Buffalo, Buffalo, NY 14260, USA
{mib,spouget}@buffalo.edu
[3] Center for Computational Research, University at Buffalo, University at Buffalo, Buffalo, NY 14260, USA.
jonesm@buffalo.edu
[4] Department of Mathematics, University at Buffalo, University at Buffalo, Buffalo, NY 14260, USA
pitman@buffalo.edu
[5] Geophysical Institute, University of Alaska, Fairbanks, AK
{pwebley,dmorton}@gi.alaska.edu

Abstract

In this paper, we present new ideas to greatly enhance the quality of uncertainty quantification in the DDDAS framework. We build on ongoing work in large scale transport of geophysical mass of volcanic origin – a danger to both land based installations and airborne vehicles. The principal new idea introduced is the concept of a localized Bayes linear model as a surrogate for the expensive simulator. Probability of ash presence is compared to earlier work.

Keywords: DDDAS, surrogates, uncertainty quantification

1 Introduction

DDDAS systems crucially depend on our ability to reliably and rapidly simulate complex systems, quantify the error and uncertainty in these simulations, dynamically observe such systems and integrate the two to improve model prediction and observation [1, 2]. The uncertainty quantification (UQ) process typically involves the use of an ensemble of simulations that sample parameter and model space. Simulations of appropriate fidelity are usually computationally expensive. A primary challenge in this paradigm then is optimizing the the ensemble since available computing resources, cost and time complexity of the simulations will limit the size of the ensemble. Towards this end, the DDDAS paradigm has promoted development of several strategies based on spectral methodologies e.g. polynomial chaos

*AFOSR Award FA9550-11-1-0336 ; PD: F. Darema

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014

and polynomial chaos quadrature with efficient quadrature schemes to minimize number of samples needed for a particular integration [3, 11] in the UQ process. An alternative idea, with a rich history of use in the statistics community, is the construction of **surrogate models called emulators using the simulation outcomes as data** and a Bayesian regression to fit the data to an appropriate surface and error model. A Gaussian process or similar construct is used for interpolation between the training points [5, 6]. Resampling this emulator will provide the UQ outcomes needed. While, the core methodology has been well established, there remain significant challenges.

We present here ideas to address these issues in the framework of the DDDAS paradigm. DDDAS brings to this domain a significant feature – the availability of data from observations to inform the simulations and the construction of the surrogates. In recent work on overground mass flows, our group has also introduced innovative methodology to surmount the computational challenges in the construction of these surrogates for UQ and in processing the very large amount of data typically encountered in 4D field simulations [8, 9].

1.1 Application

We will frame the developments in terms of a challenging application, namely, the transport of volcanic ash that we have worked on over the last 3 years. Long-range volcanic ash transport models have been given the general name Volcanic Ash Transport and Dispersion Models (VATDMs). The use of these models became well-known following the eruption of Eyjafjallajökull, Iceland, in April, 2010, when the NAME model forecasts were particularly prominent in being used as the basis for an almost complete shutdown of European air traffic. We consider in our work a VATDMs namely `puff` and `HYSPLIT` that are used by the Anchorage and Washington (DC) Volcanic Ash Advisory Center (VAAC) for real-time civil aviation forecasts of ash presence.

1.2 DDDAS Approach to Probabilistic Forecasts

We have accomplished significant goals towards our objective of using DDDAS approaches to volcanic ash forecasts. Our primary modeling tool is a new code *puffin* formed by the combination of a plume eruption model `Bent` and the ash transport model `Puff`. Data from satellite imagery, observation of vent parameters and windfields drive our simulations. Figure 1 describes the overall flow of the ash plume prediction DDDAS application. Estimates of the vent parameters and grain size distribution are sampled in a strategy outlined in our papers [3, 17]. Together with suitably chosen input numerical weather prediction ensembles, we can now create a composite ensemble of runs of the new coupled model *puffin* and Numerical Weather Prediction (NWP) choices. The weighted ensemble based outcome leads to a forecast of ash concentration for given members of the ensemble and the probability of ash presence when these are properly integrated. Satellite imagery and radiosonde data available are then used to compute a discrepancy field between observation and simulation. This discrepancy is then used to drive a feedback loop to estimate the source parameters using an estimation methodology. Ash detection based on satellite imagery is a difficult problem [15] and the disambiguation of signal into ash, water vapor cloud etc. needs a careful analysis. One of the projected outcomes of our work will be an integration of model and observation to alleviate issues rising from the difficulty of this disambiguation. The probabilistic forecast from our uncertainty analysis of the *puffin* tool is being considered as an additional source of information for feeding this disambiguation process.

The key to success in this workflow are the propagation of uncertainty – in parameters (vent radius, vent velocity, mean grain size and grain size variance) and source terms (wind field) and assimilation with data from satellites and other observations. The DDDAS components are in the systematic update of inputs, models and modeling parameters and interpretation of observations in a probabilistic framework. We have to date used an ensemble based uncertainty quantification and parameter estimation

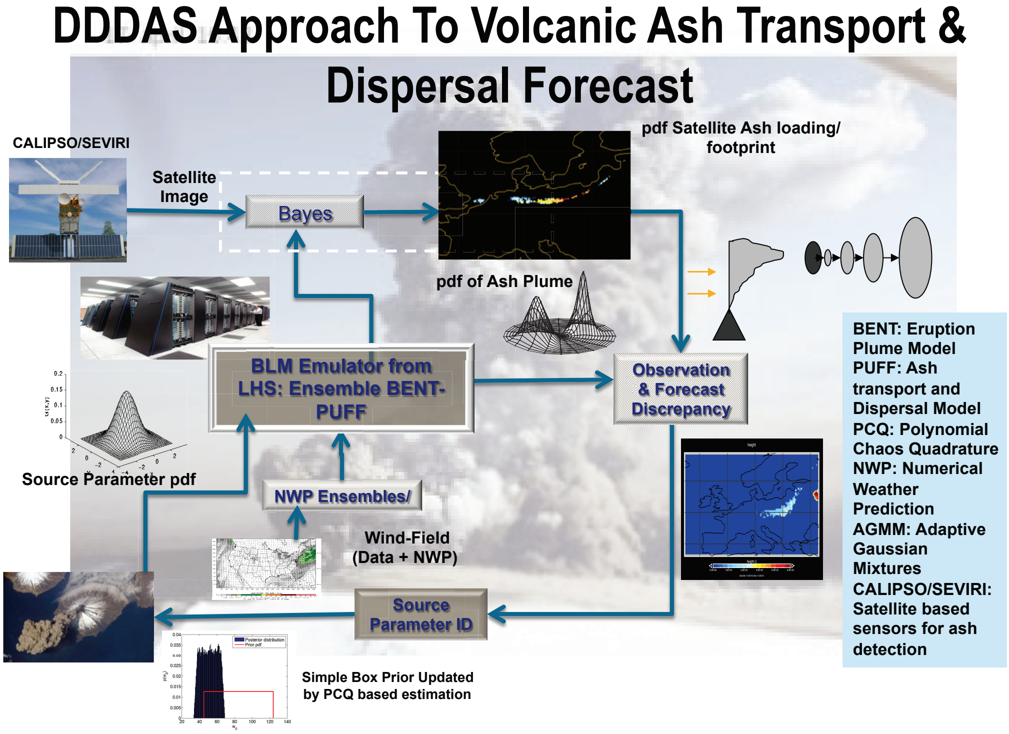

DDDAS Approach To Volcanic Ash Transport & Dispersal Forecast
CALIPSO/SEVIRI
Satellite Image
Bayes
pdf Satellite Ash loading/ footprint
pdf of Ash Plume
BLM Emulator from LHS: Ensemble BENT-PUFF
Observation & Forecast Discrepancy
Source Parameter pdf
NWP Ensembles/
Wind-Field (Data + NWP)
Source Parameter ID
Simple Box Prior Updated by PCQ based estimation
BENT: Eruption Plume Model
PUFF: Ash transport and Dispersal Model
PCQ: Polynomial Chaos Quadrature
NWP: Numerical Weather Prediction
AGMM: Adaptive Gaussian Mixtures
CALIPSO/SEVIRI: Satellite based sensors for ash detection

Figure 1: Alternate workflow of this DDDAS application as mapped to a set of heterogeneous resources and data flows. Note the replacement of the PCQ based probability calculation with a LHS design ensemble and BLM based probability calculation relative to our previous work [3].

methodology – polynomial chaos quadrature (PCQ) in combination with data integration to complete the DDDAS loop. PCQ methods suffered from an exponential growth in the number of samples needed in the ensemble when the dimensions grew. [10] developed a variation on the polynomial chaos quadrature scheme – Conjugate Unscented Transform (CUT) that addressed this issue to a large extent making possible higher dimension analysis with a quadrature type method. However, PCQ is, ultimately, a spectral method and cannot capture higher frequency behavior (in realistic time), nor easily sample from tail events from any probability distribution. The difficulty in sampling was circumvented in an *ad hoc* fashion [11] by restricting the volume inputs to only very large flows. Such a restriction is difficult to do in a systematic fashion and the predictive power of the methodology is entirely dependent on the choice of that cutoff.

To complete the DDDAS loop we assimilate of available observational data to correct and refine the model forecast helps in reducing the associated uncertainties. However, limited sensor range and sensor inaccuracies can lead to imprecise measurements. An improved solution should be a weighted mixture of model forecast and observation data. Hence,a future challenge is to design computationally tractable data assimilation tools that incorporate information from various sources while simultaneously compensating for simulation errors and observational inaccuracies. A minimum variance based approach is utilized to assimilate the model prediction with measurement data, obtained from satellite imagery. This provides us posterior statistics of uncertain parameters, given statistics of measurement data and prior statistics of model outputs which are obtained from propagation of uncertainty through dynamic model.

2 UQ using Emulators

2.1 Emulators

While, several emulators have been proposed over the years the GASP (Gaussian Process) is among the most popular. We work here with the Bayes Linear Models (BLM) that are largely equivalent in structure to GASPs and computationally more tractable. The Emulator used for our purpose attempts to fit a piecewise polynomial through already available data and estimates the mean and variance for the inputs for which simulations are not available using Bayesian Linear Regression. The moments are then adjusted using Bayes linear equations.

From Bayesian Linear Regression we have,

$$s(x) = \beta G(x) + \hat{\epsilon} \qquad \hat{\epsilon} \sim N(0, \sigma^2) \tag{1}$$

where s(x) is the response function, G is the matrix of basis functions, β is the vector of coefficients and $\hat{\epsilon}$ is the the Gaussian model of the error. The equations of Bayes Linear method which allow the adjustment of mean and variance are given by

$$E(s(y)|s(x)) = E[s(y)] + Cov[s(x), s(y)]Var[s(x)]^{-1}(s(x) - E[s(x)]) \tag{2a}$$

$$Var[s(y)|s(x)] = Var[s(y)] + Cov[s(y), s(x)]Var[s(x)]^{-1}Cov[s(x), s(y)] \tag{2b}$$

where s(x) is the simulation data which is required to adjust our belief of the response s(y), x is the collection of sample points and y is the collection of points for which simulations are not available (henceforth termed as resamples) for generating responses s(x) and s(y) respectively. E[s(y)|s(x)] is the Bayes Linear Mean and Var[s(y)|s(x)] is the Bayes Linear Variance. The functioning of emulator can be understood by superimposing equations 2 and 1 which yields

$$E(s(y)|s(x)) = g(y)\beta + r(y)^T R^{-1}(s(x) - G(x)\beta) \tag{3a}$$

$$Var[s(y)|s(x)] = \sigma^2(1 - r(y)^T R^{-1} r(y)) \, , r_i(y) = \exp\left(- \sum_{n=1}^{N_{dim}} \theta_n (y_n - x_{i,n})^2\right) \tag{3b}$$

R is the matrix of the correlation functions at x such that R$_{i,j}$ = r$_i$(x$_j$) = r$_j$(y$_i$). Equation 3a indicates that the emulator is composed of a mean which is approximated using least square fit, and an error term which is modeled as a Gaussian process with ϵ = s(x) - G(x)β being the true error evaluated at the sample points. θ_n is the vector of hyper-parameters or roughness parameters and N_{dim} is the number of dimensions associated with the data set. The emulator hyper-parameters θ_n require an expensive nonlinear optimization technique to iteratively find the optimal values [9].

With the emulator in hand, it is a simple matter to resample using multiple draws at each spatial location of interest and thereby compute a probability.

2.2 Localization and Parallelization

The computational effort in R^{-1} in (3a) dominates the workflow and for the very large data sets we have in this application the necessary repeated evaluation of this soon becomes infeasible, especially inside a DDDAS loop where the UQ must be accomplished in a short time. The key insight in making this computation feasible is a recognition of the strong locality of the correlations allowing an approximation of R^{-1} as

$$R^{-1} \approx \sum_{i=1}^{N} R_i^{-1} \text{ , } R : \Omega \to \mathbb{R}, \Omega = \cup_{i=1}^{N} \Omega_i, R|_{\Omega_i} = R_i : \Omega_i \to \mathbb{R} \tag{4}$$

The quality of the approximation is clearly important for the quality of the emulator and depends on the choice of decomposition. However, this decomposition enables us to compute an effective emulator as an approximation using an ensemble of emulators based on the decomposition of R. Each member of this ensemble uses only local data and hence can be computed in parallel! The size of each Ω_i is related to the local strength of the correlations and have to be carefully chosen. For selecting the neighborhood points Dalbey [16] proposed a tessellation based approach wherein the input parameter space is tessellated and correlations further away than 2 cells in the tessellation are disregarded. The localization and tessellation raises concern of overfitting which have been carefully addressed in earlier work [16] using a fit generated as above and tested with several hundred pieces of data outside those used in the fit. The cross validations tests in [9, 16] include comparing the BLM surface from 2048 simulations to an exhaustive evaluation of the surface with 10^4 simulations prove that this is not the case here.

While, computationally efficient, the tessellation approach is not directly related to the emulator quality. We propose here an alternate methodology based on distance metrics $D_s$in the input parameter space. A simple choice for D_s we have explored in preliminary work is $D_s = \sqrt[2]{\sum_{n=1}^{C_{dim}} (x_n - y_n)^2}$ where C$_{dim}$ is the number of dimensions considered for the euclidean distance and x and y are any two multi-dimensional points. A simple strategy for picking a local patch is to normalize each dimension to a $[0, 1]$ range and select a small distance e.g. 0.05 to pick points for the interpolation. If the number of points is insufficient for interpolation then the distance is increased by 0.005 till enough points are found.

3 Results and Discussion

In past work, based on PCQ type approaches [14] we have shown that our core methodology is sound and yields reasonably good probabilistic estimates of ash presence. Starting with the same distributions (carefully described in earlier work) we now exercise the emulator methodology. The emulator construction starts with an ensemble of 700 members of a Latin Hypercube based distribution of the 4 listed parameters (vent radius, vent velocity, mean grain size and grain size variance). Ranges and distributions for the inputs were obtained from experts. The *puffin* tool then produces ash forecast data for 700×10 elevation levels every 6 hours. We construct an emulator for each level at each time using all 700 spatially distributed data fields. A localization strategy with 30,000 of the localized emulators are then put together using an inverse distance weighted linear combination.

Figure 2 shows the output from two ensemble members of 700 which is used in making the emulator. Figure 3 shows early results from the PCQ based probability computation and the emulator based probability of ash presence. The emulator based results are somewhat different from the PCQ based results. We note that the emulator is based on a space-filling design of the input space – a Latin hypercube design. The PCQ on the other hand is based on minimizing error in moment computation.

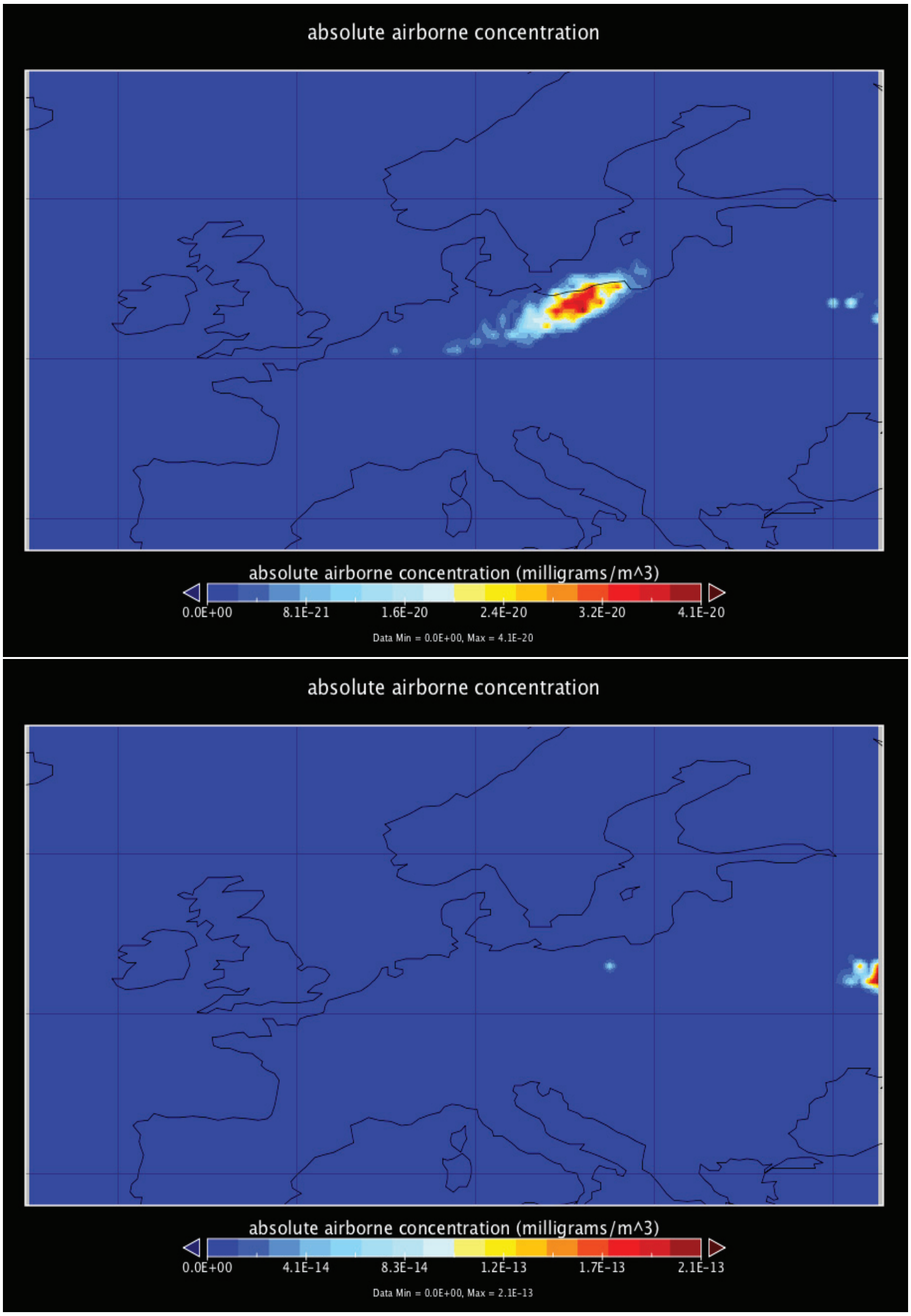
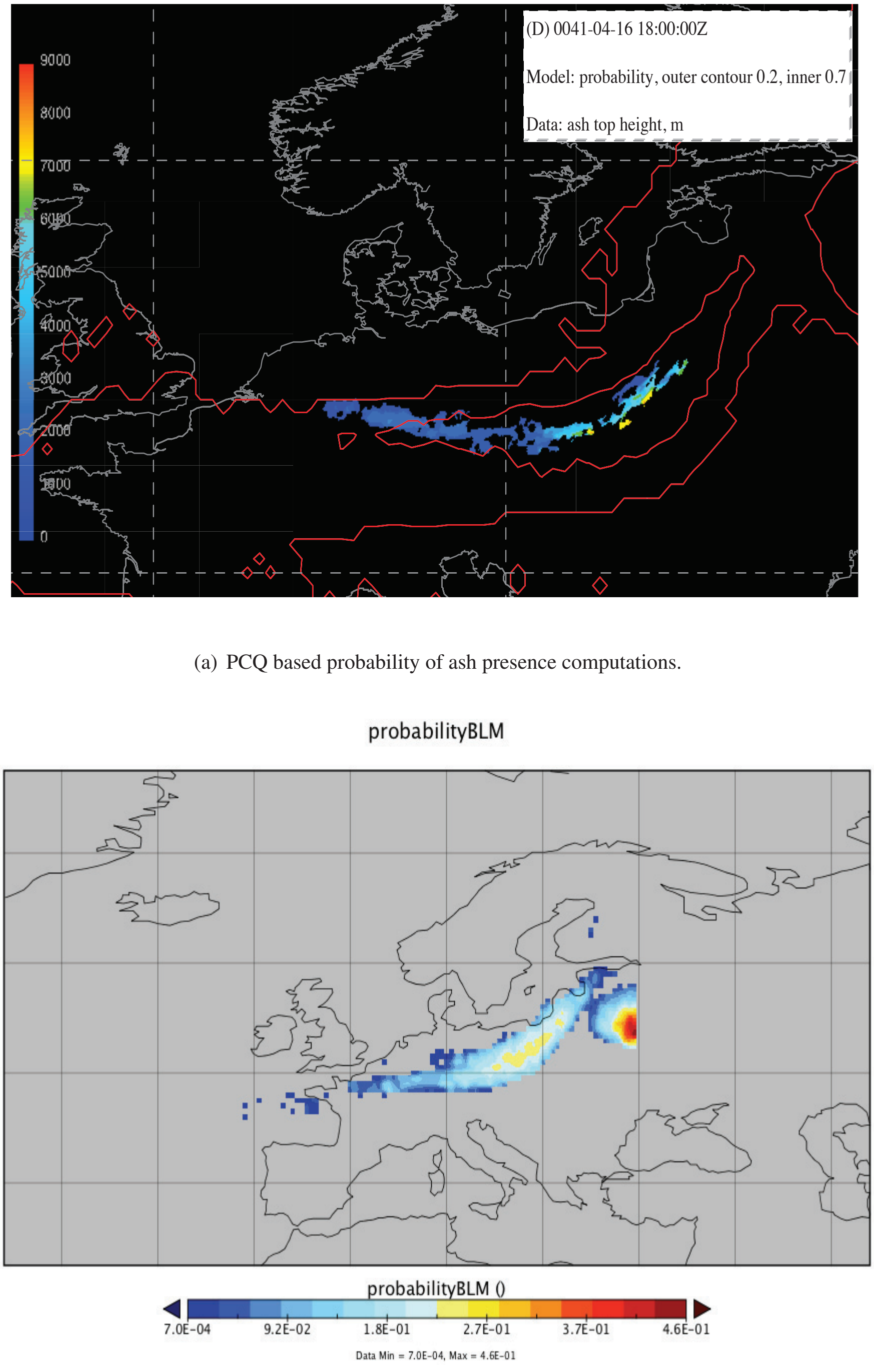

absolute airborne concentration
absolute airborne concentration (milligrams/m^3)
0.0E+00
8.1E-21
1.6E-20
2.4E-20
3.2E-20
4.1E-20
Data Min = 0.0E+00, Max = 4.1E-20
absolute airborne concentration
absolute airborne concentration (milligrams/m^3)
0.0E+00
4.1E-14
8.3E-14
1.2E-13
1.7E-13
2.1E-13
Data Min = 0.0E+00, Max = 2.1E-13

Figure 2: Puffin outcomes for two members of the ensemble of 700 used in computing the emulator.



(a) PCQ based probability of ash presence computations.

(b) BLM emulator based probability of ash presence computations.

Figure 3: a) PCQ based probability computation from [14]. Line contours 20% and 70% – solid contours are satellite observations b) The BLM emulator based probability computation for ash presence from a 700 sample ensemble. Uncertain parameters are vent diameter, vent velocity, grain size diameter and variance. While the overall footprints from both methods are similar the probabilities are different,

4 Conclusions and Future Work

In summary, we have introduced an emulator based strategy for uncertainty quantification to the propagation of ash. While, in of itself it already produces interesting and distinctly different forecasts – the key reason to consider this strategy is the possibility to a) introduce an adaptive strategy whereby ensemble members can be added and taken out based on user need (see for e.g. recent work [19]), and, b) computational advantages in terms of localization and parallelization of the processing.

References

[1] InfoSymbiotics/DDDAS – The Power of Dynamic Data Driven Application Systems, F. Darema, C. Douglas and A. Patra ed., 2010, http://www.dddas.org/afosr-nsf-workshop-2010/report/DDDAS-InfoSymbiotics%202010%20Report%20(1).pdf

[2] G. Allen, K. Baldridge, G. Biros, A. Chaturvedi, C. C. Douglas, M. Parashar, J. How, J. Saltz, E. Seidel, A. Sussman - (Editor. F. Darema): DDDAS Workshop 2006 ; in www.cise.nsf.gov/dddas

[3] R. Madankan, S. Pouget, P. Singla,M. Bursik, J. Dehn, M. Jones, A. Patra, M. Pavolonis, E. B. Pitman, T. Singh, P. Webley, Computation of Probabilistic Hazard Maps and Source Parameter Estimation For Volcanic Ash Transport and dispersion to appear J. Comp. Physics

[4] Adams, M.L. et. al. ed., Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification, Report of the COMMITTEE ON MATHEMATICAL FOUNDATIONS OF VERIFICATION, VALIDATION, AND UNCERTAINTY QUANTIFICATION, National Academies Press, 2012

[5] M. C. Kennedy, A. OHagan, and N. Higgins. Bayesian analysis of computer code outputs. In Quantitative methods for current environmental issues, pages 227243. Springer, London, 2002

[6] M. Goldstein ”Bayes Linear Analysis for Complex Physical Systems Modeled by Computer Simulators” (2012), in Uncertainty Quantification in Scientific Computing, Dienstfrey, A and Boisvert R.F. ed, Springer.

[7] Williams, B., Higdon, D., Gattiker, J., Moore, L., McKay , M. and Keller-McNulty, S. (2006) Combining Experimental Data and Computer Simulations, With an Application to Flyer Plate Experiments. Journal of Bayesian Analysis 1, Number 4, Sep 28, pp 765–792, DOI:10.1214/06-BA125

[8] R. Shivaswamy, A. Patra, V. Chaudhary, Integrating Data and Compute Intensive Workflows for Uncertainty Quantification in Large Scale Simulation Application to Model Based Hazard Analysis, to appear International Journal of Computer Mathematics

[9] E. R. Stefanescu, M. I. Bursik, A. K. Patra, Effect of digital elevation model on Mohr-Coulomb geophysical flow model output, Natural Hazards (2012) DOI 10.1007/s11069-012-0103-y.

[10] A. Nagavenkat, P. Singla, T. Singh, Conjugate unscented transform rules for uniform probability density functions, Proceedings of the 526 American Control Conference, 2013

[11] K. Dalbey, A.K. Patra, E. B. Pitman, M. I. Bursik, and M. F. Sheridan (2008), Input uncertainty propagation methods and hazard mapping of geophysical mass flows, J. Geophys. Res., 113,B05203, doi:10.1029/2006JB004471

[12] M.J. Bayarri, J.O. Berger, E. Calder, K. Dalbey, S. Lunagomez, A.K. Patra, E.B. Pitman, E. Spiller, R.L. Wolpert, Using Statistical and Computer Models to Quantify Volcanic Hazards, Technometrics, Vol 51, 2009

[13] M. J. Bayarri, J. O. Berger, E. S. Calder, A. K. Patra, E. B. Pitman, E. T. Spiller, R. L. Wolpert, A Methodology for Quantifying Volcanic Hazards, in review for SIAM Review.

[14] A. Patra, M. Bursik, J. Dehne, M. Jones, M. Pavolonis, E. B. Pitman, T. Singh, P. Singla, P. Webley, A DDDAS Framework for Volcanic Ash Propagation and Hazard Analysis, DDDAS Workshop, ICCS 2012, Omaha, NE, June 2012.

[15] M. Pavolonis, A. Heidinger, J. Sieglaff, Automated Retrievals of Volcanic Ash and Dust Cloud Properties from Upwelling Infrared Measurements, to appear Journal of Geophysical Research.

[16] K. Dalbey, Predictive simulation and model based hazard maps of geophysical mass flows, PhD Dissertation, University at Buffalo, 2009.

[17] M. Bursik, M. Jones, C. Carn, K. Dean, A. Patra, M. Pavolonis, E. Pitman, T. Singh, P. Singla, P. Webley, H. Bjornsson, M. Ripepe, Estimation and propagation of volcanic source parameter uncertainty in an ash transport and dispersal model: application to the Eyjaf- jallajokull plume of 14 - 16 April 2010, Bull Volcanoll.

[18] F. Liu and M. West, A dynamic modelling strategy for Bayesian computer model emulation, Bayesian Analysis (Impact Factor: 2.42). 01/2004; 1. DOI:10.1214/09-BA415

[19] E., Bayarri, M., Berger, J., Calder, E., Patra, A., Pitman, E., and Wolpert, R., Automating Emulator Construction for Geophysical Hazard Maps Spiller, SIAM/ASA Journal on Uncertainty Quantification 2014 2:1, 126-152