

Inversion and forward estimation with process-based models: an investigation into cost functions, uncertainty-based weights and model-data fusion

Maricar Rabonza^{a,b,*}, Michele Nguyen^b, Sébastien Biass^c, Susanna Jenkins^{a,b}, Benoit Taisne^{a,b} and David Lallemand^{a,b}

^a*Earth Observatory of Singapore, Nanyang Technological University, 639798, Singapore*

^b*Asian School of the Environment, Nanyang Technological University, 639798, Singapore*

^c*Department of Earth Sciences, University of Geneva, Geneva, 1211, Switzerland*

ARTICLE INFO

Keywords:

Tephra
Inversion
Model calibration
Kriging
Kelud volcano

ABSTRACT

Process-based models for inversion and forward estimation have supported our understanding of earth, environment, and hazards processes. These methods are often applied on spatial data without accounting for their spatial nature and uncertainty. Using an example of reconstructing past volcanic eruption characteristics and associated tephra fallout from different sets of field observation, we demonstrate the importance of making the best use of data-related uncertainty and spatial information in inversion and forward estimation. We present strategies for: (1) the selection of appropriate cost functions accounting for their behaviour and implied distribution of residuals, (2) the treatment of differential uncertainty when combining multiple data, and (3) the leveraging of both model and data when estimating the spatial distribution of output. Results show that a data-informed choice of cost function and accounting for uncertainty and spatial characteristics of data leads to consistent improvements in model predictive performance for both inversion and forward models.

1. Introduction

Process-based models are important tools used to represent physical phenomena and processes in natural hazards analysis. In the *inversion or calibration setting*, they are used to estimate parameters that cannot be directly measured, through calibration to the observed response of the hazard system. Examples include estimating floodplain characteristics based on observations of past floods, estimating earthquake rupture characteristics based on recordings from seismic stations, or estimating volcanic eruption source parameters based on measurements of the deposited tephra (ejected particles of all sizes) across the impacted region (Li et al., 2022; Georgoudas et al., 2007; Connor and Connor, 2006). In the *forward estimation setting*, process-based models are used to estimate the *response* of the hazard system based on calibrated or assumed parameters. Examples include flood simulation using a calibrated inundation model, estimating ground-motion intensity for a given earthquake magnitude and location, or estimating tephra accumulation given eruption and wind parameters (Uhlenbrook et al., 2004; Worden et al., 2010; Wang et al., 2022; Hurst and Turner, 1999; Folch et al., 2009). Such applications of inversion and forward estimation frameworks merge the imperfect knowledge derived from data with that of process-based models in a way that utilises the reinforcing characteristics of both models and data.

As attention turns to such learning-from-data problems, recent advancements in optimisation and uncertainty quantification underline the issue of treating inversion-forward estimation problems with a black-box perspective that ignores the characteristics of both models and data (Hollós et al., 2022; Cabaneros and Hughes, 2022; Biegler et al., 2011; Isaac et al., 2015; Oden et al., 2010). Despite technological improvements in the collection of observed data (e.g. remote sensing) and the increasing complexity of newly-developed process-based models, inversion-forward estimation frameworks often ignore particular characteristics in the data that they rely on (e.g. distribution, uncertainty, and spatial properties; Willcox et al. (2021)). Thus, there is a need to be strategic about how the inversion-forward estimation framework uses limited and uncertain data.

*Corresponding author

✉ maricar1001@ntu.edu.sg (M. Rabonza); (M. Nguyen); (S. Biass); (S. Jenkins); (B. Taisne); (D. Lallemand)

ORCID(s):

In this paper, we study this issue using the example of reconstructing past volcanic eruption characteristics and associated tephra fallout from observations across the impacted area. The application follows a typical inversion-forward estimation workflow shown in Figure 1. The inversion and forward estimation workflows can be conducted sequentially, as visualized in Figure 1, or separately. In the inversion/calibration setting, tephra fallout models are used to estimate eruption source parameters (ESPs, e.g. tephra mass, total grain size distribution, plume height, and plume profile) and empirical parameters (e.g. fall-time threshold, diffusion coefficient) from observations of tephra characteristics such as deposit thickness and grain size, typically collected through field surveys in the impacted area. As in any inversion process, an automated search routine finds the parameters of a model in such a way that the modelled response best approximates the observed data. The search routine optimises a *cost function* (also known as loss function or objective function), which calculates the agreement between the model estimates of tephra deposits and the observations. This process can estimate eruption parameters for past events, which is an important input for estimating the potential size and intensity of future eruptions, and their associated consequences (Newhall and Self, 1982; Carey and Sparks, 1986; Pieri and Baloga, 1986; Armienti et al., 1988; Scarpati et al., 1993; Mastin et al., 2009; Stohl et al., 2011; Pouget et al., 2013; Madankan et al., 2014; Bear-Crozier et al., 2020). In the forward estimation setting, tephra fallout models can map the distribution of tephra accumulation. The map can serve multiple purposes, including estimating eruption impacts on communities (Le Pennec et al., 2012; Wardman et al., 2012; Magill et al., 2013; Biass et al., 2017), assessing the vulnerability of buildings to tephra loading (Williams et al., 2020; Hayes et al., 2019; Spence et al., 2005), or quantifying risk to agriculture (Gómez-Romero et al., 2006; Ayras and Delmelle, 2012; Thompson et al., 2017).

Given the importance and broad applicability of process-based models, this study focuses not on the models themselves, but on how they are calibrated with limited and uncertain data and how they can best utilise such data in both inversion/calibration and forward estimation settings. We explore these through the study of tephra fallout from the 2014 eruption of Kelud volcano and the use of the *Tephra2* model of Bonadonna et al. (2005) as the process-based model. Connor and Connor (2006)'s algorithm is followed for the inversion workflow. The study does not make any particular claim or conclusions about the characteristics of the Kelud eruption or the accuracy of the *Tephra2*, both of which have been extensively studied by others (Maeno et al., 2019; Caudron et al., 2015; Hargie et al., 2019; Suzuki and Iguchi, 2019; Connor et al., 2011; Connor and Connor, 2006). Instead, this study focuses on methodological contributions into:

1. The choice of cost functions in the calibration process: What is the impact of this choice? What are the implied assumptions linked with this choice?
2. Making use of multiple data sources with varying uncertainty: How can we benefit from all data while still accounting for varying uncertainty?
3. Combining model estimates and data in the forward-model spatial estimate: How can we make use of both model results and observed data to estimate and map tephra accumulation? How do we account for varying uncertainty of the data in this estimation?

We believe that the recommendations can benefit researchers interested in improving their estimates when conducting inversion and estimation of spatially-distributed data. The approach to select the cost function, treat uncertain data, and generate spatial estimates can also be applied to other earth and environmental models. The paper proceeds as follows. Section 2 explains the current state of the art and limitations of interest in using process-based models for inversion-forward estimation of tephra fallout. Section 3 introduces our test case: modelling the tephra fallout from the 2014 eruption of the Kelud volcano using *Tephra2*. The Methods section (Section 4) illustrate the proposed improvements for model calibration and spatial estimation. We provide a combined Results and Discussion in Section 5 and the conclusions in Section 6.

2. Current approach and limitations

Tephra2 is an accessible and popular process-based model for inversion (Connor and Connor, 2006) and estimation of tephra fallout used in many volcanology studies (e.g. Volentik et al. (2010); Costa et al. (2012); Bonadonna and Costa (2013); Mannen (2014); Bonadonna et al. (2015); Connor et al. (2019); Constantinescu et al. (2021); Williams et al. (2020)). It solves the advection-diffusion equation analytically using wind and eruptive conditions as inputs. The model output is the mass accumulation and grain-size distribution of tephra deposits.

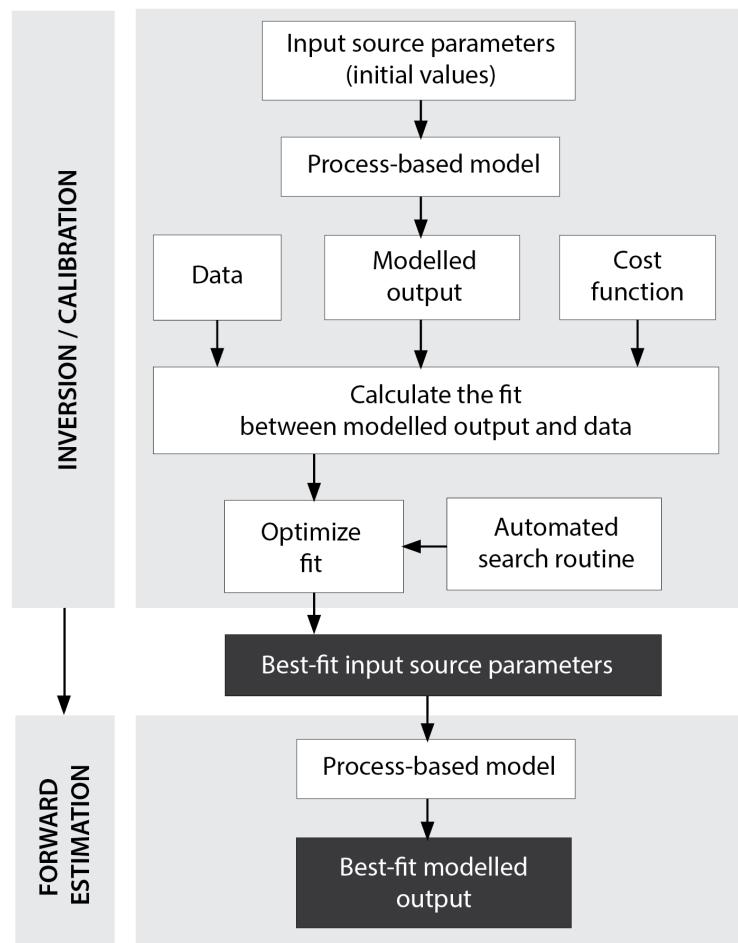


Figure 1: Schematic of a typical inversion-forward estimation workflow using a process-based model. The purpose of the inversion is to estimate the best-fit input source parameters (e.g. erupted mass, plume height), while the purpose of forward estimation is to obtain best-fit modelled outputs (e.g. a map of the tephra thickness). The inversion and estimation workflows can be conducted sequentially, as visualised here, or conducted separately/independently. The outputs of the inversion and estimation are shown as black boxes.

2.1. Choice of cost function

The process of fitting *Tephra2* to data involves minimizing a cost function. The goal of the cost function is to mathematically define how much a modelled output deviates from the observation, i.e. goodness of agreement or the model fit (see, e.g., Thorne and Stephenson (2001)). In the current version of the *Tephra2* inversion code, three possible cost functions are available to the user for inversion:

$$\text{Mean square error (MSE)} = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2, \quad (1)$$

$$\text{Chi square error} = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2 / x_i, \quad (2)$$

$$\text{Tokyo log error} = \sum_{i=1}^n \left(\log \frac{y_i}{x_i} \right)^2, \quad \text{where } \log \frac{y_i}{x_i} = 0, \text{ if } \frac{y_i}{x_i} \leq 0 \quad (3)$$

where y_i is the modelled output, x_i is the observation, and n is the total number of data points.

In the paper by Connor and Connor (2006) where the Tephra2 inversion algorithm is first presented, the chi-square cost function was used. Connor and Connor (2006) mentioned that this cost function allows thin and thick deposit measurements to be treated equally in the optimisation. Other than in this work, there are no further instances in literature where the choice of the cost function is discussed (Volentik et al., 2010; Costa et al., 2012; Bonadonna and Costa, 2013; Mannen, 2014; Bonadonna et al., 2015; Connor et al., 2019; Constantinescu et al., 2021; Scollo et al., 2008; Fontijn et al., 2012).

Outside the field of tephra fallout modelling, multiple studies have brought attention to the fact that different measures of model performance (implied in the choice of cost-function) may satisfy different desirable characteristics (Chen et al., 2017; Makridakis, 1993; Armstrong and Fildes, 1995). Certain cost functions penalize different magnitudes and directions of forecast error differently (Walther and Moore, 2005). An underestimation may not have the same penalty as an overestimation (Morley et al., 2018). Some characteristics may also be relevant for pragmatic purposes such as their interpretability. Such considerations help in narrowing down desirable cost functions applicable to an application.

Beyond the cost functions' characteristics, the takeaway from these studies is that no metric is inherently better for all applications. Importantly, choosing a cost function implies making an assumption on the type of distribution of the residuals (where residuals are the difference between the modeled output and the observations) (Engle, 1993). Hence, in its correct application based on the assumed residual distribution, a cost function is optimal. For instance, Hodson (2022) presented a theoretical justification that root mean square error is optimal for normal (Gaussian) residuals while mean absolute error (MAE) is optimal for Laplacian residuals. In Section 4.2, we implement this knowledge of the cost functions' characteristics and inherent assumptions on the residuals to demonstrate how an optimal cost function may be selected for a test case. We investigate the extent that the cost function affects the resulting estimates of tephra fallout and the associated residuals.

2.2. Treatment of varying uncertainty in data

For the same deposit, several factors may contribute to varying uncertainty between data points depending on how, when and where measurements are collected (Engwell et al., 2013; Bonadonna et al., 2015). According to Engwell et al. (2013), there are two types of uncertainties in tephra thickness. Uncertainty can be due to natural variation, which is related to the physical process of deposition, preservation, and remobilisation. The second type is observational uncertainty, or those uncertainties related to differing measurement techniques. Measurements that are most reliable and contain the least uncertainty are those taken from a well-preserved deposit, i.e. those taken soon after an eruption has ended in areas with little deposit reworking (Pyle, 2016; Blong et al., 2017). However, such conditions are often difficult to meet even for recent eruptions, and impossible when studying old eruptions. Large portions of tephra deposits may also be inaccessible while they are still well-preserved (Walker and Croasdale, 1971). Field campaigns conducted at a significantly later time after the eruption might acquire measurements subjected to post-eruption processes (e.g., compaction, soil formation, bioturbation, and remobilisation (Engwell et al., 2013)) or local weather conditions (e.g., rain or wind (Hayes et al., 2002; Wilson et al., 2011; Arnalds et al., 2013; Blong et al., 2017; Oishi et al., 2018; Dominguez et al., 2020)). Therefore, uncertainty in measured data is rarely quantified, or even reported in field studies and in literature. Likewise, in the context of inversion-forward estimation, uncertainty in data is rarely accounted for. Given limited data, it is advantageous to make use of all available data, but the proper treatment of uncertainty becomes ever more important when using multiple data sources with differing levels of uncertainty.

In the current *Tephra2* inversion algorithm, all data are treated equally in the optimisation step. Relative differences in the uncertainty between datasets, when ignored, could influence the inversion and tephra load estimates. To investigate the extent of this influence, we present a calibration approach in Section 4.3 that weighs the observations in the cost function based on the reliability of the data. Our analysis highlights the importance of accounting for such uncertainties in fitting the model to the data. We note that this approach is developed not to find or optimise for the most appropriate weight for the measurements. Quantifying the absolute uncertainties in the data is also beyond the scope of the work.

2.3. Making use of model and data

The issues and limitations presented in Sections 2.1 and 2.2 relate to the inversion setting, the procedure which enables us to estimate source parameters to characterise the eruption. In the following, we look at the forward estimation setting in which the goal is obtaining tephra fallout estimates across a large spatial domain. The current process of forward estimation, as visualised in Figure 1 involves two key components: best-fit input source parameters and

the process-based model. The model takes in the best parameter values and produces model estimates. While they should represent the best fit to the model in terms of the chosen cost function, the modelled outputs may diverge from observations in a spatially structured way due to e.g., model approximations, unaccounted processes, and uncertainties inherent in any model. While these model-data disagreements may not cause issues for applications such as tephra volume estimation when spatial aggregates are used, they affect forward forecast or prediction performance when spatially explicit estimates are the focus and the process-based models are unable to capture the spatial complexity one might observe in the field. Hence, if the goal of the forward estimation is to obtain the spatial predictions that best agree with observations, the process-based model alone may not serve as the best tool.

In Section 4.4, we present a spatial estimation method which harnesses both the process-based model and the spatial structure of the data. The proposed method aims to improve estimation by combining model and data, a concept similar to the goal of data assimilation methods in weather forecasting. In addition to treating the tephra fallout data as spatial and accounting for their spatial characteristics, the estimation method considers observations as imperfect versions of the true process, with uncertainties always associated with them. In line with the varying data uncertainties mentioned in Section 2.2, we investigate two variations of the estimation method. One approach weighs all the data points equally, while the other accounts for the relative uncertainty between different groups of data.

3. Test case: 2014 Kelud eruption

3.1. Eruption characteristics

The 2014 eruption of Kelud volcano, located in East Java, Indonesia was selected as a test case for this study. The explosive activity started at 22:50 local time on 13 February 2014 and lasted for over four hours (Global Volcanism Program, 2014). The eruptive activity finally declined on 14–17 February 2014. The total erupted volume was estimated to be $0.25\text{--}0.50 \text{ km}^3$ (bulk deposit volume, $0.14\text{--}0.28 \text{ km}^3$ in dense rock equivalent), and the mass eruption rate was $6.5 \pm 2.8 \times 10^7 \text{ kg/s}$ (Maeno et al., 2019). The impacts of tephra fallout were widespread with over 76,000 people evacuated, forty regional flights cancelled and rerouted, and more than 26,000 buildings destroyed or damaged (Global Volcanism Program, 2014; Williams et al., 2020; IFRC, 2014).

A documentation of the eruption sequence by Maeno et al. (2019) and numerical simulations by Tanaka et al. (2016) highlighted that the dispersal process and tephra accumulation were affected by the local winds. At high altitudes of $\sim 17 \text{ km}$ above sea level, the umbrella cloud was affected by strong winds from the east. Low altitudes of $\sim 5 \text{ km}$ above sea level were affected by winds from the southwest, transporting tephra to the northeast and causing a bilobate tephra deposit. Published isopachs of accumulated tephra for this eruption (Figure 2) show a subtle bilobate feature that is attributed to the dynamics and evolution of the eruption and the local winds around the volcano (Maeno et al., 2019).

3.2. Tephra fall data

The study utilises two sets of tephra fall data, each derived from field surveys by different teams at various times after the eruption. The analysis takes the data in terms of load. When only thickness measurements are available, they are converted to loads based on a deposit bulk density of 1400 kg/m^3 measured by Maeno et al. (2019). The first dataset, here called *Dataset 1*, is the tephra load data obtained from a field study of thickness by Universitas Gadjah Mada (UGM) (Anggorowati and Harijoko, 2015) and later used in inversion modelling by Williams et al. (2020). The measurements were taken 2–3 days post-eruption at 81 locations within 2 to 60 km of the vent (Williams et al., 2020). The other data set, here called *Dataset 2*, consists of tephra load data converted from thickness measurements from a geological survey by Maeno et al. (2019) conducted a month after the eruption. Dataset 2 provides load information at 50 locations, including areas up to 75 km north of the vent that Dataset 1 doesn't cover. For this study, we exclude the six eyewitness reports also presented by Maeno et al. (2019).

Four outliers were removed from the raw datasets as they were an order of magnitude different than nearby data, potentially related to issues of data collection or significant deposit reworking. Being outside the range of the other observations, these were deemed to have a disproportionately strong influence over the model fits. After removing these outliers, the dataset used in our analyses consists of 127 points. A map of the datasets is presented in Figure 3.

3.3. Previous inversion study

The value of using inversion and forward estimation with process-based models for Kelud volcano has been presented in recent literature. Williams et al. (2020) applied inversion as one of the methods to support remote assessment of tephra fall building damage and vulnerability assessment of buildings around Kelud volcano. In Williams

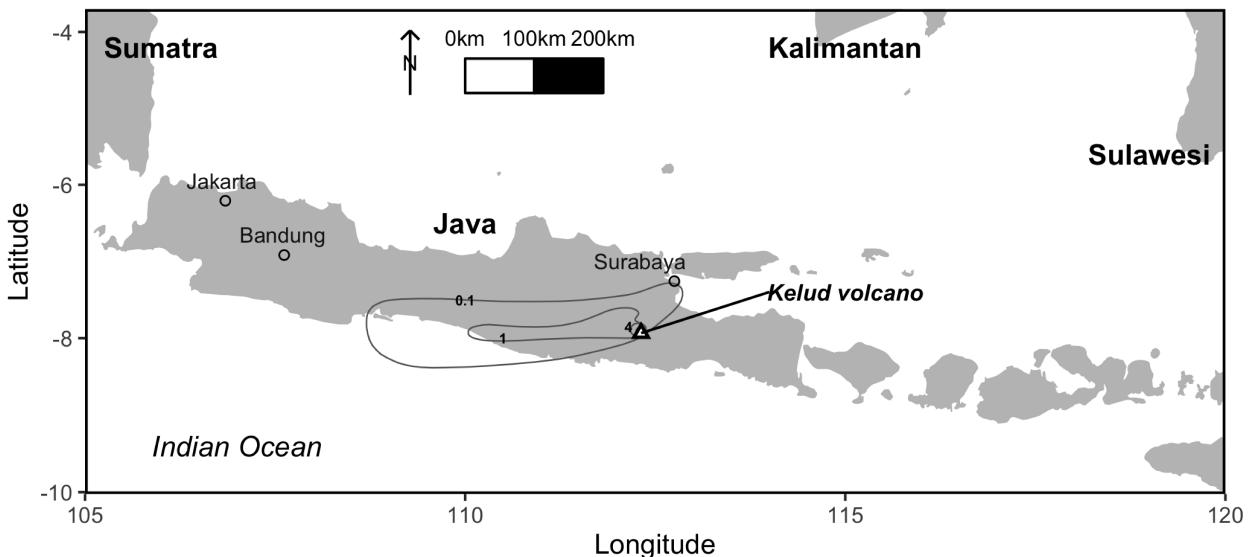


Figure 2: Location map of Kelud volcano in Indonesia. The contours represent tephra thicknesses of 0.1, 1, and 4 cm adopted from Maeno et al. (2019)'s study of the 2014 Kelud eruption fallout deposit.

et al. (2020)'s study, thickness measurements were inverted to estimate the best-fit source parameters, which were later used to map a continuous deposit. Such application is important to enhance our knowledge of risk to tephra hazards on the scale of buildings or regions, especially when there is a high likelihood of future damaging eruptions from Kelud (Maeno et al., 2019).

4. Methods

4.1. Model setup for *Tephra2* inversion

Tephra2 produces tephra load estimates based on approximations of physical equations for dispersion. The inversions use fixed wind conditions as model inputs. These are expressed as a single wind profile based on the wind dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) Era-Interim Reanalysis for the midnight of 13 February 2014 (Dee et al., 2011).

The parameter bounds used in the inversion algorithm are shown in Table 1. The ranges of values are selected based on studies in the literature of the 2014 Kelud eruption and a recent inversion for the same eruption in Williams et al. (2020). For this study, some of the ranges of values are widened relative to Williams et al. (2020)'s to ensure that the inversion solution converges within the limits set. We use the default optimisation routine from Connor and Connor (2006)'s inversion algorithm, the Nelder-Mead simplex method.

4.2. Evaluation of cost function

In this paper, we consider five cost functions. They consist of functions already in the current *Tephra2* code (i.e., mean square error written as Equation 1) and chi-square error as in Equation 2), as well as three additional functions:

$$\text{Mean absolute error (MAE)} = \frac{1}{n} \sum_{i=1}^n |y_i - x_i| \quad (4)$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - x_i}{x_i} \right| \quad (5)$$

$$\text{Mean square log error (MSLE)} = \frac{1}{n} \sum_{i=1}^n \left(\log \frac{(y_i + 1)}{(x_i + 1)} \right)^2 \quad (6)$$

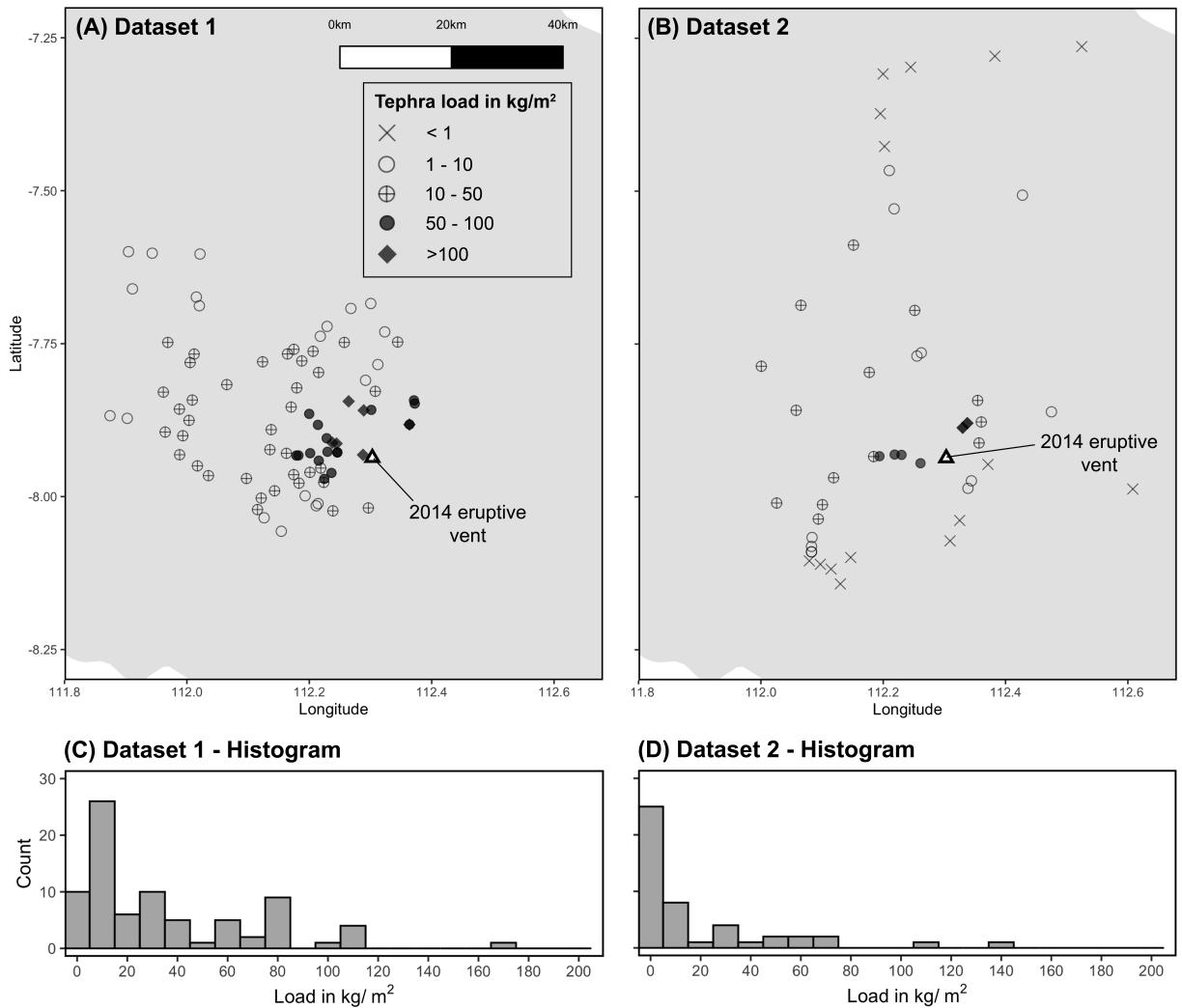


Figure 3: Map and histogram of datasets used in the study. The location of Kelud volcano's vent from the 2014 eruption is based on Goode et al. (2019)'s study of the eruption.

where y_i refers to the model estimate, x_i is the observation, and n is the total number of data points.

Instead of Tokyo log error (Equation 3), this study uses MSLE (Equation 6), a more well-known cost function in statistics and machine learning. In cases where the observed and predicted values are strictly positive, as in the case of tephra load data, the Tokyo log function is essentially the same as MSLE. In MSLE's formula, the '+1's attached to the observed value x_i and predicted value y_i ensure mathematical stability since $\log(0)$ is undefined and both y and x can be 0.

We propose a two-step approach to evaluate the choice of cost function for the *Tephra2* inversion. First, the choice should be guided by the properties of the cost function. For instance, some cost functions are more sensitive to outliers, and some are better for data which span orders of magnitude. The properties of the cost functions are summarised in Table 2 and discussed in Section 4.2.1, which serves as a guide to narrow down the appropriate cost functions for the test case.

The second step is to run test inversions using the shortlisted cost functions and the set up in Section 4.1, and analyse the resulting residuals. While the choice of the cost function depends on the properties most relevant to the context, it also implies an assumption about the distribution of residuals. MSE, for instance, assumes normally-distributed

Table 1

The inversion in this study uses the range of initial values presented in this table. The initial values are selected based on information from: [a] Kristiansen et al. (2015), [b] Maeno et al. (2019), [c] Goode et al. (2019); Maeno et al. (2019), [d] Goode et al. (2019); Maeno et al. (2019), and [e] Williams et al. (2020).

Inverted parameters	Values from literature	Initial values for the inversion
Plume height (km above sea level)	18 to 26 [a]	15 to 23
Erupted mass 10^{10} kg	38 to 66 [b]	10 to 100
Median grain size (phi)	-3 to 1 [c]	-3 to 2.5
Phi standard deviation	1 to 3 [d]	0.5 to 3
Fall-time threshold (s)	0.001 - 10,000 [e]	0 to 15,000
Diffusion coefficient ($m s^{-2}$)	0.0001 - 10,000 [e]	0 to 15,000
Plume profile (α)	3 [e]	3
Plume profile (β)	0.001 - 3 [e]	0.001 to 3.5

residuals. Since each cost function is associated with distributional assumptions on the residuals, we can check the suitability of the choice of cost function using goodness-of-fit tests described in Section 4.2.2.

4.2.1. Selection based on cost functions' properties

The foundation of any cost function is the model residual, defined as $\varepsilon = y - x$, where x refers to the observed value and y is the predicted value of the model. Cost functions almost always include a transformation of the observed value, predicted value or the residual. This transformation influences the model estimates according to these properties: (1) order-dependence, (2) sensitivity to outliers, and (3) symmetry. The equations for the transformed residuals associated to the cost functions are shown in Table 2.

Order-dependence describes how the function covers orders of magnitude for the residual. For instance, an order-dependent cost function may utilise the relative error ε/x (e.g. in the chi-square cost function) or the log-transform of the observed and predicted values (e.g. in mean square log error, MSLE). On the other hand, a cost function that's not order-dependent, also known as scale-dependent functions, keep the residual ε as is in the function to retain the units/scale of the observed and modelled values. A comprehensive description of scale-dependent cost functions is provided in Walther and Moore (2005).

If the modeller wishes to balance the treatment of distal and proximal deposits, order-dependent cost functions may be desirable. Since order-dependent cost functions are scaled to the magnitude of the measured value, observations made in thin parts of the deposit are equally important as observations made in the thicker parts of the deposit. This aspect is important due to a few reasons. Tephra deposits thin exponentially with distance from the vent, and outcrops mapped across a single deposit may span orders of magnitude (Pyle, 1989). Distal measurement of tephra accumulations are important to inform the extrapolation of the thinning rate beyond the outermost isopach. Examples of common order-dependent cost functions include chi-square, MSLE and MAPE.

Some cost functions are more sensitive to outliers than others. Cost functions in which the residual is squared, for instance, penalise large residuals more heavily than small residuals. Common examples are MSE and root-mean-square-error (RMSE), both of which minimise to the same solution. If we prefer to constrain penalty on large errors and make the function more resistant to outliers, the absolute residual $|\varepsilon|$ can be used instead of ε^2 , for instance, with the mean absolute error (MAE). MAE may be more appropriate for instances when the residuals are not normally distributed, when large model residuals don't have to be weighted heavily, and when the presence of outliers is a significant issue.

Lastly, the symmetry of the cost function relates to the treatment of underestimation versus overestimation. MSE, MAE, and chi-square treat residuals symmetrically. MSLE applies more penalty for underestimation than overestimation. Mean absolute percentage error (MAPE) penalises overestimation more than underestimation.

4.2.2. Selection based on cost functions' assumption on residuals

The choice of cost function also implies a distribution of residuals. So in addition to their theoretical properties, we can examine the validity of the cost functions by investigating the distribution of resulting residuals using statistical tests such as the Kolmogorov-Smirnov (K-S), the Cramér-von-Mises (CvM) and the Anderson-Darling (A-D) as well as the Shapiro-Wilk (S-W) (Ramachandran and Tsokos, 2015; Stephens, 1986). These tests use different metrics or test

Table 2

Cost functions with their formulae, residual transformation and assumed error distributions. In the formulae, x refers to the observation, y the model estimate, n the number of data points and $\epsilon = y - x$ the residual. For the Gaussian distributions, $\hat{\sigma}$ denotes the estimated standard deviation, while for the Laplace distributions, b denotes the estimated scale. Characteristics of the cost functions are provided with a recommendation of when these are useful for tephra load inversion. The checkmarks indicate if the cost function satisfies a specific characteristic.

	Cost function				
	MSE	MAE	Chi-square	MSLE	MAPE
Formula	$\frac{1}{n} \sum_{i=1}^n \epsilon_i^2$	$\frac{1}{n} \sum_{i=1}^n \epsilon_i $	$\frac{1}{n} \sum_{i=1}^n (\epsilon_i^2 / x_i)$	$\frac{1}{n} \sum_{i=1}^n \left(\log \frac{(y_i+1)}{(x_i+1)} \right)^2$	$\frac{100}{n} \sum_{i=1}^n \left \frac{\epsilon_i}{x_i} \right $
Transformed residual Z_i	<i>Predicted – Actual</i>	<i>Predicted – Actual</i>	$\frac{\text{Predicted} - \text{Actual}}{\sqrt{\text{Actual}}}$	$\log_{10} \left(\frac{\text{Predicted} + 1}{\text{Actual} + 1} \right)$	$\frac{\text{Predicted} - \text{Actual}}{\text{Actual}}$
Assumed distribution	$\frac{Z_i}{\hat{\sigma}} \sim N(0, 1)$	$\frac{Z_i}{b} \sim \text{Laplace}(0, 1)$	$\frac{Z_i}{\hat{\sigma}} \sim N(0, 1)$	$\frac{Z_i}{\hat{\sigma}} \sim N(0, 1)$	$\frac{Z_i}{b} \sim \text{Laplace}(0, 1)$
Scale-dependent	The units of the observed and predicted values are retained. Useful for interpretability .				
Order-dependent	Thinner deposits are as important as thicker deposits. Useful when spanning orders of magnitude.				
Sensitive to outliers	Estimates fit large values closely. Useful when these are the most reliable and informative.				
Robust against outliers	Reduced penalty on large residuals. Useful when there are many outliers in the data.				
Symmetric	Treats overestimation and underestimation equally. Useful to avoid anomalous skewness in errors.				
Asymmetric	Treats overestimation and underestimation unequally. Useful for avoiding either scenario.				
				\checkmark^*	\checkmark^{**}

*Larger penalty for underestimation. **Much larger penalty for overestimation.

statistics to measure how similar the empirical distribution of the residuals is to its assumed one. For example, the K-S test statistic is:

$$D = \max(|F_0(\epsilon) - F_n(\epsilon)|), \quad (7)$$

where $F_0(\epsilon)$ is the assumed cumulative distribution function and $F_n(\epsilon)$ is the empirical cumulative distribution of the residuals. The larger D is, the more unlikely the residuals to have been generated from the assumed distribution. Goodness-of-fit can also be assessed graphically using quantile-quantile (Q-Q) plots which compare the empirical quantiles from the model fit to the theoretical quantiles from their assumed distributions. While both statistical tests and Q-Q plots are useful indicators of goodness-of-fit, they both assume that the residuals are themselves independent and identically distributed.

4.3. Weighting the data in the cost function

In order to account for different levels of uncertainty inherent in our datasets, we propose to weigh each data point based on its uncertainty relative to a reference dataset. In this way, less reliable data have correspondingly less influence on the inversion. The reference dataset may be an individual measurement or a group of measurements having the least measurement uncertainty. In this work, we use the prior information that Dataset 2 has a relatively larger uncertainty than Dataset 1 because of the delay in field data acquisition between the two sets of data. Thus, we set Dataset 1 as the reference dataset. In the inversion algorithm, points in this reference dataset are assigned a weight of 1. Other datasets are then assigned weights relative to the uncertainty of the reference data. For example, the points from a dataset assumed to be twice as uncertain as the reference dataset may be assigned a weight, w , of 0.5.

Using weights based on relative uncertainty leads to a weighted cost function, where each data point is weighted relative to the reference dataset. With our proposed weighting scheme, the weighted MSE cost function, for instance, can be written as:

$$\frac{1}{n} \sum_{i=1}^n w_i (y_i - x_i)^2 \quad (8)$$

where w_i indicates the uncertainty weight assigned to the data point i . For our application we set Dataset 2's weight to 0.5.

We evaluate the effectiveness of the weighted cost function based on the accuracy of the forward estimates on unseen (out-of-sample) data. The test set consists of a random subset (20%) of Dataset 1, while the training set consists of the rest of the Dataset 1 points and all of Dataset 2 (Figure 4). A test set should ideally consist of data that closely resembles the true deposit, so a stratified test set was selected to represent thin and thick deposits in relatively equal proportion as the true deposit. Inversion is conducted using the training points, the parameter ranges in Table 1, and the weighted form of the cost function. For all the inversions, we generate forward estimates of tephra load on the locations of the test points. To assess predictive performance, we compare the test errors between the unweighted inversion and weighted inversion. The test errors are calculated following the formula of the cost function used in the inversion.

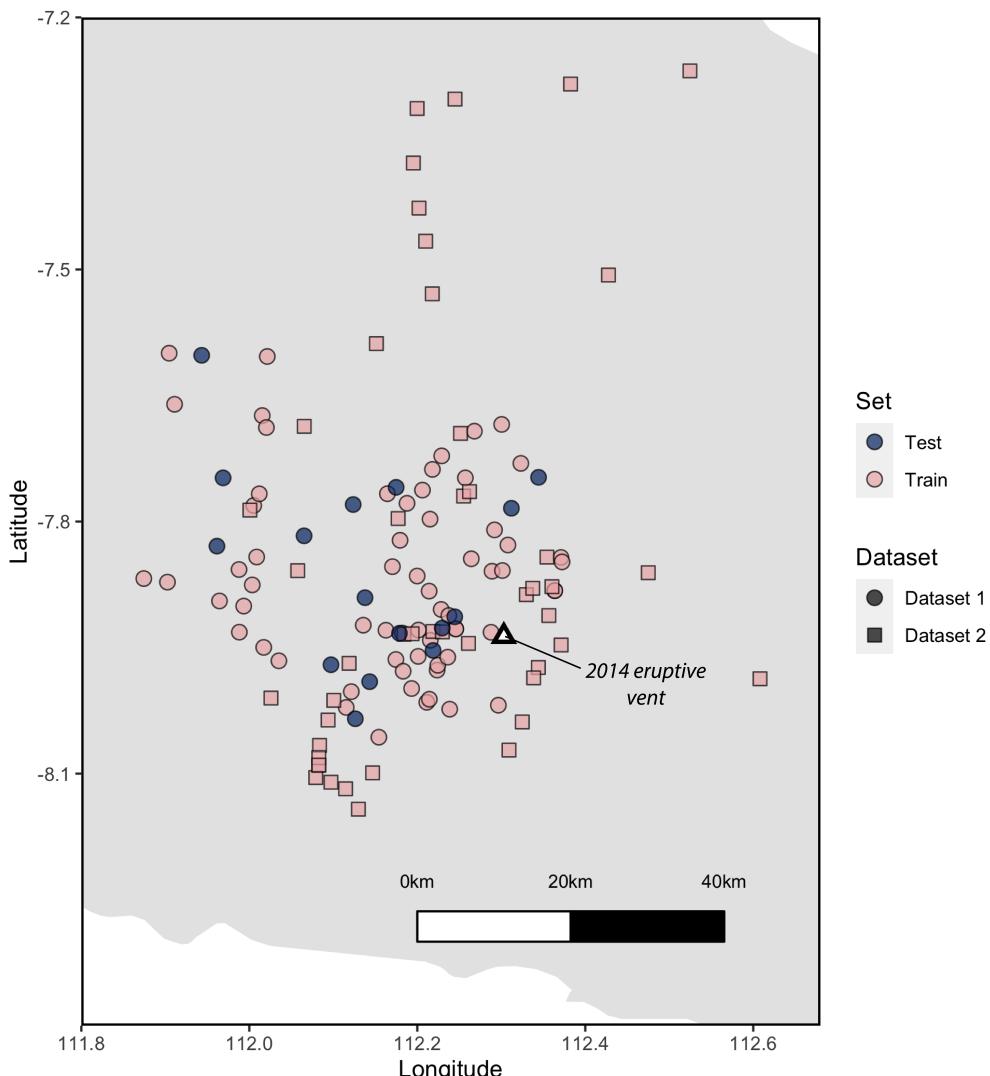


Figure 4: The effectiveness of the weighted cost function was evaluated using a train-test split process. Shown is a map of the training and test points used for the methods in Section 4.3.

4.4. Combining model estimates and data

To estimate the spatial distribution of tephra load across the entire area, we propose to combine both forward model estimates as well as the data. The process consists of three general steps: (1) calculate the residuals between the forward model and data, (2) interpolate these residuals using spatial statistics techniques, and (3) combine the interpolated residuals with the forward model. Before performing the interpolation, the model residuals can be transformed to an

appropriate form as suggested by any of the formulas for Z_i in Table 2. We test and present two different spatial statistics techniques for the interpolation step. The result is an improved estimate of tephra load that combines the information of the model and data. We elaborate on the process in this section.

Our methods use kriging, one of many interpolation techniques that estimate values at unobserved locations using a limited set of known spatial data. In kriging, the interpolation accounts for the spatial arrangement of the data in such a way that points nearby the site of interest are given more weight than those farther away. This approach differs from another popular method, inverse distance weighted interpolation, in that we do not assume a spatial distribution model beforehand but estimate it from data. Another advantage of kriging is that we can estimate the uncertainty associated with the interpolated value.

4.4.1. Unweighted model-data fusion

We describe the procedure for an *unweighted* model-data fusion approach via kriging. The approach is unweighted as it doesn't account for any differential uncertainty between individual or groups of data. In the methods that follow, we adopt simple kriging, which assumes that the difference between model and data all over the study area has a mean of zero and follow a multivariate normal distribution (Wackernagel, 2003). Performing kriging involves choosing a *variogram model* and checking the adequacy of the assumptions by the model. The variogram model defines the spatial relationship between each pair of points in the data. Mathematically, it relates the degree of variation in the sampled values and the spatial distance between these points. The parameters of the variogram model are typically user-defined or fitted with statistical software that uses approaches like maximum likelihood and least-squares estimation. Typically, we assume that the variogram and hence the distribution of the residuals is stationary and isotropic. This means that the characteristics of the distribution, particularly how the correlation decreases with pairwise distance, do not change with the location or direction.

1. Calculate residuals Z_i (from Table 2) between the best-fit modelled output and data:
 - (a) Run the inversion following Section 4.1 to find the best-fit parameters. For the calibration of parameters, use the most appropriate cost function identified using the steps in Section 4.2.
 - (b) Using *Tephra2* and the best-fit parameters, estimate the load of tephra fall at all sites (sampled and unsampled).
2. Interpolate the residuals using kriging:
 - (a) For the sampled sites, calculate the residuals using the equations for Z_i in Table 2 that corresponds to the cost function used in Step 1a.
 - (b) Conduct a simple kriging interpolation to predict the residuals at all sites (sampled and unsampled sites) in the study area. The residual at any prediction location s_o can be calculated as:

$$\hat{Z}(s_o) = \sum_{i=1}^N \lambda_i Z(s_i) \quad (9)$$

where $Z(s_i)$ is the residual at an observed location s_i , λ_i is the kriging weight applied to the value at the observed location, and N is the number of observed sites. The kriging weights λ_i are calculated from a variogram model and a covariance matrix Σ , which describes the spatial relationships among the values at the observed sites and the prediction location. The Appendix (A.3) provides a step-by-step procedure for simple kriging interpolation that includes formulas for λ_i and the covariance matrix Σ .

3. Combine the best-fit modelled output and the interpolated residuals:
 - (a) Back-transform the kriging predictions $\hat{Z}(s_o)$ to produce residuals with the same units as the data (See the Appendix A.4 for sample formulas to use when back-transforming the kriging predictions).
 - (b) For all locations in the study area, add the estimates of tephra load from Step 1b to the back-transformed residuals from Step 3a. This produces an updated map of tephra fallout that accounts for both the model and the observations.

4.4.2. Weighted model-data fusion

In practice, we could have different amounts of uncertainty associated with different datasets collected, as we do in our case study due to different survey teams measuring tephra load at different time points after the eruption. To

account for varying data uncertainty in the interpolation procedure presented in Section 4.4.1, we can use an extension of simple kriging introduced by Worden et al. (2018), which they use in the context of earthquake ground motion intensity estimation. The method starts by identifying a *reference dataset* - a subset of the data with least uncertainty. Then, we define the additional uncertainty associated with other datasets relative to this reference dataset.

For the Kelud test case, we set Dataset 1 as the reference dataset because it was acquired the earliest after the eruption. We can set an uncertainty weight of $w_{k=1} = 1$ for Dataset 1 and $w_{k=2} = 0.5$ for Dataset 2. More generally, for datasets other than the reference dataset (Dataset $k > 1$), we will always have $w_k < 1$. Based on these uncertainty weights, we can define an adjustment factor that relates the variance of the reference dataset to other datasets (Dataset $k > 1$) as $a_k = \sqrt{w_k}$.

The procedure follows the same steps as in Section 4.4.1 except for Step 2b. With an adjustment factor, the equation for the residual at any prediction location (originally Equation 9) would become:

$$\hat{Z}(s_o) = \sum_{i=1}^N \lambda'_i a_k Z(s_{i,k}) \quad (10)$$

where $Z(s_{i,k})$ is the residual at observation site i and associated with dataset k . The adjusted kriging weight, λ'_i , is calculated based on an adjusted covariance matrix Σ' . The adjusted covariance matrix can be written as:

$$\Sigma' = \Omega \odot \Sigma \quad (11)$$

where $\Omega = \mathbf{a}\mathbf{a}^T$, \mathbf{a} is the vector of adjustment factors corresponding to the observation sites s_1, \dots, s_N , and \odot denotes the element-wise multiplication. Note that the covariance matrix and variogram model parameters were estimated using the reference dataset only.

Similar to the simple kriging approach in Section 4.4.1, we can estimate the uncertainty associated with the interpolated values. Also following Steps 3a and 3b in Section 4.4.1, we can add the interpolated residuals to the optimised model estimates. The result is an updated map of the tephra load that accounts for the model estimates, the observations, and the varying uncertainty between different sets of data.

4.4.3. Comparison of the kriging methods

We assess the performance of the kriging methods in Section 4.4.1 and 4.4.2 using leave-one-out cross-validation (LOOCV). Cross-validation is a popular method to compare the test performance between models. The total number of points used for testing is maximised via iterative sampling and model fitting. LOOCV is a type of cross-validation where just a single observation is held out for testing in every iteration; thus estimating the out-of-sample estimation performance (Shao, 1993).

We can also check if the kriging interpolation overfits the training points. To confirm this, we compare the LOOCV errors to pre-kriging training errors calculated using the same performance metrics (root mean square error, chi-square, and mean square log error) from the process-based model fit. Since in LOOCV we are conducting out-of-sample estimation, LOOCV errors which are higher than the training errors from the *Tephra2* fit would suggest that kriging overfits the data.

5. Results and Discussion

Our study shows that there is a significant influence on the inversion results and forward spatial estimates due to (1) the choice of the cost function in inversion, (2) the treatment and accounting of differential uncertainty of the data, and (3) the treatment and accounting of the spatial characteristics of the data. In the following subsections, we detail these in turn.

5.1. Use of a suitable cost function

In order to identify the characteristics of cost functions most applicable to the test case, we assessed the distribution of data values and the presence of outliers. The data cover a wide range of values (near zero to 168 kg/m^2 , see Figures 3C and 3D) so we select an order-dependent cost function for a balanced treatment of data across different orders of magnitude. We also decided to use a cost function that is not sensitive to outliers because of the presence of relatively few large values located close to the vent. Using Table 2 as a guide, the cost functions that satisfy such characteristics are MSLE and chi-square.

Table 3

Goodness-of-fit tests for different cost functions. The null hypotheses for the Kolmogorov-Smirnov (K-S), the Cramér-von-Mises (CvM), the Anderson-Darling (A-D) and the Shapiro-Wilk (S-W) tests are that the residuals follow the assumed distributions. So, large p-values indicate better adherence to the assumptions. For our tephra case study, the residuals from the MSLE model fit seem to fit its assumed distribution the best. The Shapiro-Wilk test is only applicable for Gaussian distributions, thus, inapplicable for cost functions with underlying Laplace distributions, such as MAE and MAPE.

Cost function	K-S statistic (p-value)	CvM statistic (p-value)	AD statistic (p-value)	S-W statistic (p-value)
MSE	0.155 (0.082)	0.285 (0.149)	1.684 (0.138)	0.937 (0.003)
chi-square	0.216 (0.004)	0.966 (0.003)	4.728 (0.004)	0.948 (0.009)
MSLE	0.060 (0.967)	0.039 (0.938)	0.327 (0.916)	0.916 (0.205)
MAE	0.156 (0.079)	0.243 (0.198)	1.520 (0.172)	not applicable
MAPE	0.245 (0.001)	1.264 (0.001)	6.003 (0.001)	not applicable

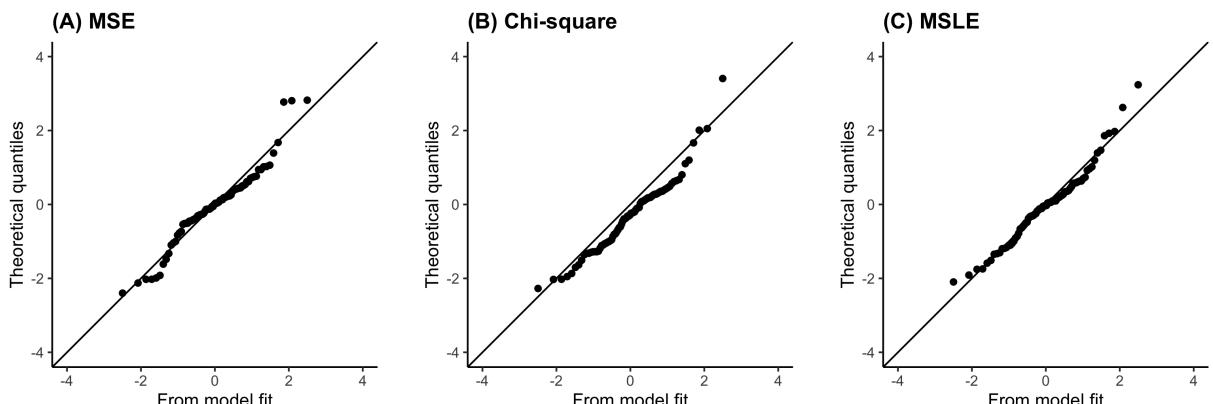


Figure 5: Quantile-quantile (Q-Q) plots illustrating goodness-of-fit for the models fitted using different cost functions: MSE, chi-square loss and MSLE. The fit is evaluated by how well the empirical quantiles from the transformed model residuals line up with the theoretical quantiles from the assumed distributions along the diagonal line.

Although MAPE is also order-dependent, and most interpretable among all the cost functions in Table 2 (because the errors it produces are in terms of percentages), we identify it as inappropriate for the test case for reasons beyond its sensitivity to outliers. The Statistics literature has provided evidence of its numerous weaknesses as a cost function. When MAPE is used for optimisation, estimated values become consistently low (Tofallis, 2015). Small actual values (less than one) result to extremely large MAPEs, while zero actual values yields infinite errors (Kim and Kim, 2016). Numerous small data are observed in the test case, which is a common occurrence in tephra fallout data. We also provide a short summary of issues associated with MAPE in the Supplementary Information.

Between MSLE and chi-square, we identify MSLE as the more suitable cost function because the residuals associated to the inversion with MSLE best adhered to the cost function's statistical assumption. This was demonstrated with MSLE giving the largest p-values across all the goodness-of-fit tests in Table 3. The Q-Q plots in Figure 5 also support the conclusion because the matched theoretical and empirical quantiles lie along the diagonal line for MSLE. The Q-Q plots for the cost functions with underlying Laplace distributions (MAE and MAPE) are given in the Supplementary Information.

In summary, while little attention is usually given to the choice of the cost function in the inversion of source parameters, our analyses highlight that the selection of cost function needs to be a conscious choice since each has associated assumptions on how the model would fit the data. Results in Figure 5 show that using different cost functions

can have a strong influence on the residuals of the best-fit modelled output. More importantly, residual analysis can be used to draw conclusions on the appropriateness of the selected cost function. Using the methods presented, we found that MSLE performs well and has characteristics well-suited for the case study.

5.2. Weighting the cost function based on varying uncertainty

The purpose of adding uncertainty-based weights to the cost function in the inversion is to fit the model to the more reliable data, while still considering the information provided by other less reliable data. In the test case, Dataset 2 is the less reliable dataset compared to Dataset 1; nevertheless, Dataset 2 is important to consider as it provides information in the deposit that might not be captured by Dataset 1. For instance, only Dataset 2 consists of load data that are less than 1 kg/m^2 . These relatively small measurements are located in areas north and south of the vent where Dataset 1 does not cover (Figure 3).

The improved fit to the more reliable data can be visualised by looking at how the distribution of residuals change when using a weighted inversion versus those of an unweighted inversion. For the test case's weighted inversion, the residuals at the locations of the reference dataset (Dataset 1) became more concentrated around zero compared to those of the unweighted inversion (Figure 6A). This difference in the distributions in Figure 6A indicate that the weighted inversion was successful in terms of fitting better to the reference dataset. As expected, the distribution of residuals for Dataset 2 from the weighted inversion does not indicate an improved fit (Figure 6B). The residual distribution for Dataset 2 only show a subtle increase in positive residuals when the inversion was weighted.

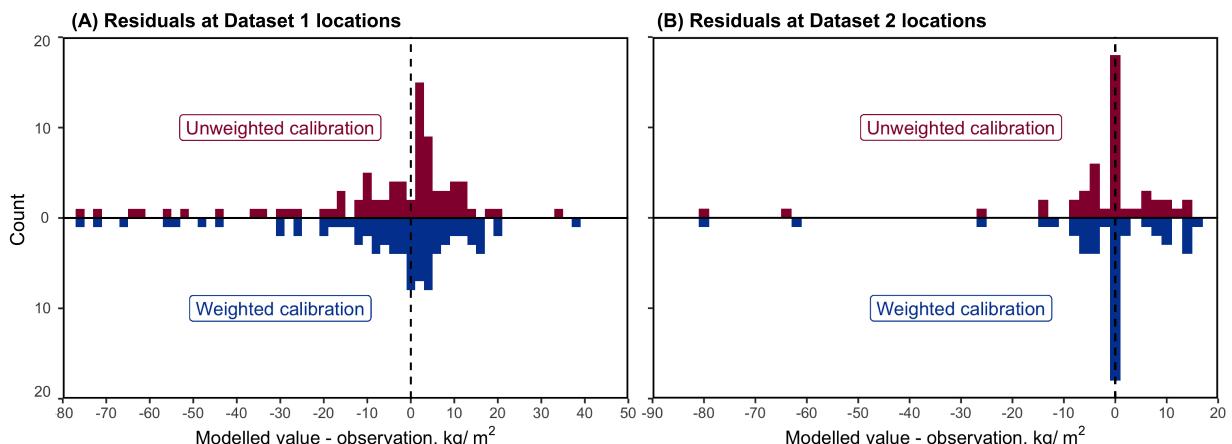


Figure 6: A visualisation of improving the fit to the more reliable dataset (Dataset 1) through a weighted inversion. Shown are histograms of residuals for inversions that make use of an unweighted and weighted MSLE cost function. Shown in (A) are residuals at locations of Dataset 1, whereas (B) shows the residuals at Dataset 2 locations.

We assessed how uncertainty-based weights in the cost function impact the best-fit modelled output of an inversion-forward estimation analysis for tephra fallout. Following the method in Section 4.3, we conducted the inversion using MSLE, the cost function that worked best based on the results in Section 5.1. Since the selected cost function was MSLE, the test errors were calculated using the MSLE formula (Equation 6). When no weights were applied to the training points, the test error is 0.05, whereas, with weights applied, the test error is 0.03 (Table 4). Using weights in the inversion resulted in a lower test error, which implies that using weights resulted in a better predictive performance at out-of-sample locations. We checked whether using different cost functions would result in the same behaviour. Thus, we repeated the inversions using the same train-test split, uncertainty weights, and initial parameter ranges but using cost functions other than MSLE. The resulting test errors, shown in Table 4, indicate that using weights in the inversion consistently reduces test errors, and thus improves the prediction capability of the inversion-forward estimation analysis.

5.3. Influence on the source parameters

Thus far, the presented methodological contributions on the inversion (use of appropriate cost functions, and applying weights to the cost function) were investigated in terms of their impacts on the best-fit modelled outputs

Table 4

Multiple inversions were conducted using the same training points (see train-test split in Figure 4) and initial parameter ranges (Table 1), but using different cost functions. The first set of inversions was run with no uncertainty-based weights applied to the training points. Shown in column (a) are the resulting errors at the test points calculated in the units of the corresponding cost function used in the calibration. Column (b), on the other hand, has weights applied to the training points in the inversion. The test errors show that applying weights results in a consistent decrease in the test errors leading to better predictive performance.

Cost function used in the calibration	Test errors	
	(a) No weights applied to training points	(b) Weights applied to training points
MSE	123.4	109.6
MAE	8.77	8.75
Chi-square	106.1	97.5
MSLE	0.05	0.03
MAPE	68.5	67.1

Table 5

Best-fit parameter values from inversions using different cost functions. All inversions used unweighted cost functions, Dataset 1 for training, and the range of parameter values in Table 1 (asl = above sea level).

Cost function (no weights applied to training points)	Inverted parameters' best-fit value							
	Column height (km asl)	Erupted mass (10^{10} kg)	Grain size Median	Grain size standard deviation	Fall time threshold (s)	Diffusion coefficient (m s ⁻²)	Plume profile, α	Plume profile, β
MSE	23.3	87.7	1.24	1.37	6356	10221	3	3.38
MAE	22.3	66.5	0.79	0.58	14380	10288	3	3.38
Chi-square	23.6	89.9	1.87	2.15	1406	12116	3	2.92
MSLE	17.4	67.7	0.85	1.06	2070	14748	3	1.30
MAPE	24.0	57.8	1.03	1.27	4937	5389	3	2.98

and the associated residuals. Acknowledging that the primary focus of an inversion is to optimise for the best-fit source parameters, we summarise the best-fit parameter values from inversions using different cost functions in Table 5 and for different weighting schemes in Table 6. The purpose of the paper is to not make claims about the true source parameters for the test case, instead we highlight that the source parameters are sensitive to the cost function and weighting in the inversion.

In Section 5.1, we discuss how the order-dependence and outlier-sensitivity properties of cost functions may influence the model fit to large (often proximal to the vent) and small values (often distal). In Section 5.2, the weighted approach allows a better fit to the more reliable data, no matter the data's magnitude. The relationship of the source parameters to these characteristics, i.e. the estimation of thicker deposits near the vent or thinner distal deposits, is affected by the physics of tephra transport. For instance, a larger plume height may lead to more tephra deposition at distal sites due to tephra being dispersed farther downwind, while a lower plume height may lead to thicker proximal deposit because of less wind advection. The relationship of observations made in distal and proximal sites to source parameters such as the erupted mass and plume height parameters have been studied by Suzuki et al. (1983), Bonadonna et al. (2005), and Yang et al. (2021).

A full understanding of the sensitivity of the proposed methods to each source parameter would require more extensive sensitivity analyses and preferably a test case with a bigger dataset. Parameter sampling algorithms, such as Markov Chain Monte Carlo algorithms, are popular for evaluating the benefits of new inversion methods (e.g. White et al. (2017); Yang et al. (2021)). Such methods may run the inversion thousands of times using different starting seeds to produce a range of fitted parameter values. The algorithms can extract confidence intervals for the source parameters, which can represent the impact of the new methods on the optimised parameters. While such sensitivity analysis is out of scope of this paper, our analysis highlights the importance of cost function selection and weighted inversion when varying data uncertainties are present when such studies are implemented. In this way, our work can contribute towards improving our understanding of relationships between data uncertainty and source parameters in the inversion.

Table 6

Best-fit parameter values from inversions using different cost functions. All inversions make use of Dataset 1 and 2 for training. The range of parameter values in Table 1 were used to initiate the inversions. The weighted cost functions follow the procedure in Section 4.3. (asl = above sea level)

Cost function and associated weighting scheme	Inverted parameters' best-fit value							
	Column height (km asl)	Erupted mass (10^{10} kg)	Grain size Median (ϕ)	Grain size standard deviation (ϕ)	Fall time threshold (s)	Diffusion coefficient ($m s^{-2}$)	Plume profile, α	Plume profile, β
MSE								
Unweighted	23.0	104.4	1.96	1.81	1639	7596	3	2.82
Weighted	23.4	85.3	1.43	1.78	1810	10936	3	3.15
MAE								
Unweighted	24.0	64.7	0.98	1.04	5647	5490	3	3.35
Weighted	23.3	91.9	1.62	1.50	4939	5129	3	3.13
Chi-square								
Unweighted	23.7	47.5	0.56	0.83	14507	3456	3	2.94
Weighted	23.4	49.0	0.63	0.90	13490	3612	3	2.85
MSLE								
Unweighted	19.5	100	1.97	1.87	4846	9019	3	1.39
Weighted	16.6	76.2	1.31	1.55	13993	10657	3	0.96
MAPE								
Unweighted	17.2	16.0	0.25	1.26	14476	2835	3	0.37
Weighted	23.5	44.6	0.46	0.64	14852	1971	3	2.72

5.4. Model-data fusion

We provided a methodology in Section 4.4 to combine observations and model estimates in such a way that the relative uncertainties between data are accounted for. The methodology increases the value of the information from the data by using them to examine the spatial structure in the model residuals which can be used together with the information from the model to fill gaps at unsampled locations. Here, we present an analysis of the new estimates of tephra load and associated residuals produced by fusion methods. One advantage of using kriging in the fusion methods is the ability to produce not only the kriging estimates, but also the uncertainties associated to the new estimates. We provide a map of uncertainties related to the new estimates of tephra load in this section in the Supplementary Information, Figure A.4.

The manually-contoured isopachs for the 2014 Kelud event by Maeno et al. (2019) (Figure 2) show a bilobate shape as influenced by the wind conditions during the tephra dispersion. However, this two-lobe pattern was not captured by *Tephra2* without the additional spatial estimation using its residuals (Figure 7). Shown in Figure 7 are tephra deposit maps prior and after implementing an unweighted fusion approach described in Section 4.4.1 when MSLE was chosen as the cost function. We see that the model-data fusion approach results in tephra load contours with a two-lobed feature, in line with the manually-contoured isopachs, and the wind shear that occurred across the eruption column,.

The maps in Figure 7 also illustrate how the measurements influence the resulting tephra deposit after the fusion. The fusion approach (Figure 7B) accounted for Dataset 2's small measurements southeast of the vent (load values ranging from less than 1 to 10 kg/m^2 in Figure 3) resulting in lower values of load in this area compared to those in Figure 7A. Due to the influence of thick deposit measurements near the vent, the fusion approach produced higher peak load values ($\sim 140 \text{ kg/m}^2$) compared to those in the best-fit modelled output ($\sim 100 \text{ kg/m}^2$), possibly reflecting sedimentation from the column's edge. Implementing a weighted fusion approach (Section 4.4.2) to combine the model in Figure 7A and data results to the tephra deposit map in Figure 9B.

While it is difficult to compare the results from the unweighted and weighted fusion based on the contour maps, we can look at the LOOCV to compare their spatial estimation performance. To compare the two approaches fairly, we evaluate estimation performance using metrics corresponding to the cost function assumed. The results in Table 7 show that both weighted and unweighted model-data fusion approaches lead to smaller LOOCV errors, and they are smallest for the weighted fusion approach. The same result is seen when other correlation structures such as exponential and Gaussian functions are used instead of the Matérn variogram (see Supplementary Information, Table A.1). This highlights the advantage of making use of data in the spatial estimation, and the importance of accounting for data

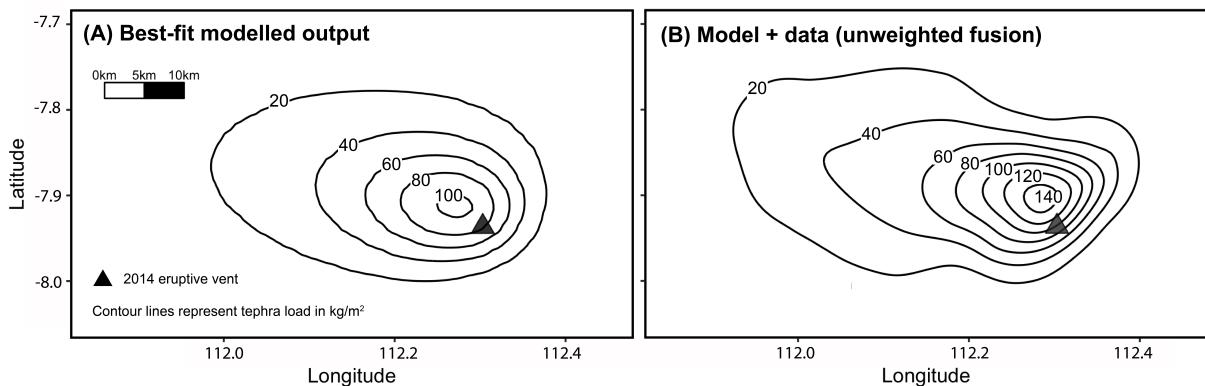


Figure 7: The map in A shows the best-fit modelled tephra deposit from an inversion with MSLE as the cost function and with weights applied based on the approach in Section 4.3. Shown in B is the result of implementing an unweighted fusion (Section 4.4.1) of the model in A and data. The fusion approach makes use of both Datasets 1 and 2, of which some are located beyond the bounds of the maps shown.

Table 7

Measures of performance of the unweighted and weighted kriging-based fusion methods in terms of LOOCV errors. For each cost function choice, the errors were calculated using the corresponding performance metric: root mean squared errors for the MSE cost function, chi-square errors for chi-square cost function and mean square log error for MSLE. The training errors before applying the fusion methods are consistently higher than the LOOCV errors, which indicates no overfitting in the kriging method. The Matérn variogram model was used to define the spatial correlation structure. For the formulas shown, x refers to the observed value, y is the model estimation, and n is the number of data points.

Cost function	Performance metric	Formula	Pre-kriging training error	LOOCV error after unweighted fusion	LOOCV error after weighted kriging
MSE	Root-mean-square	$\sqrt{\frac{1}{n} \sum_{i=1}^n \epsilon_i^2}$	16.166	15.470	15.064
Chi-square	Chi-square	$\frac{1}{n} \sum_{i=1}^n (\epsilon_i^2 / x_i)$	8.051	5.362	5.300
MSLE	Mean square log	$\frac{1}{n} \sum_{i=1}^n \left(\log \frac{(y_i+1)}{(x_i+1)} \right)^2$	0.0570	0.040	0.033

uncertainties for generating spatial estimates. Comparing the LOOCV errors to pre-kriging training errors also confirms that the approach is not overfitting.

We further examine the residuals after weighted and unweighted fusion. Figure 8 shows the histograms of these residuals at the locations of Datasets 1 and 2 separately when MSLE is used for inversion and fusion. We observe that using the weighted fusion approach reduces the spread/variability of the residuals of Dataset 1 points, indicating improved fit. This is expected since the weighted approach prioritises the fit to the more reliable dataset. On the other hand, the weighted fusion results to a wider spread in the residuals of Dataset 2, which shows the influence of applying less weight.

The importance placed on the fitted process-based model and the data can be balanced by the *nugget parameter* in the variogram model of the kriging procedures. The nugget determines the spatial variability of the modelled quantities in short distances and hence, affects the smoothness of the predicted residual surface: the larger the nugget, the smoother the surface. This variation in smoothness in relation to the nugget value is visualised in Figure 9. Here, we present different spatial models of varying nuggets and the associated map of tephra fallout for the MSLE case.

The fusion methods use and build upon simple kriging, which comes with an assumption that the mean of the residuals is zero. Alternatively, if the spatial factors underlying the residuals are known, regression kriging may be used instead of simple kriging. In addition to the variogram that describes the stochastic spatial relationship between residuals, we could use spatial covariates such as topography or distance from the vent to model their mean surface (trend model).

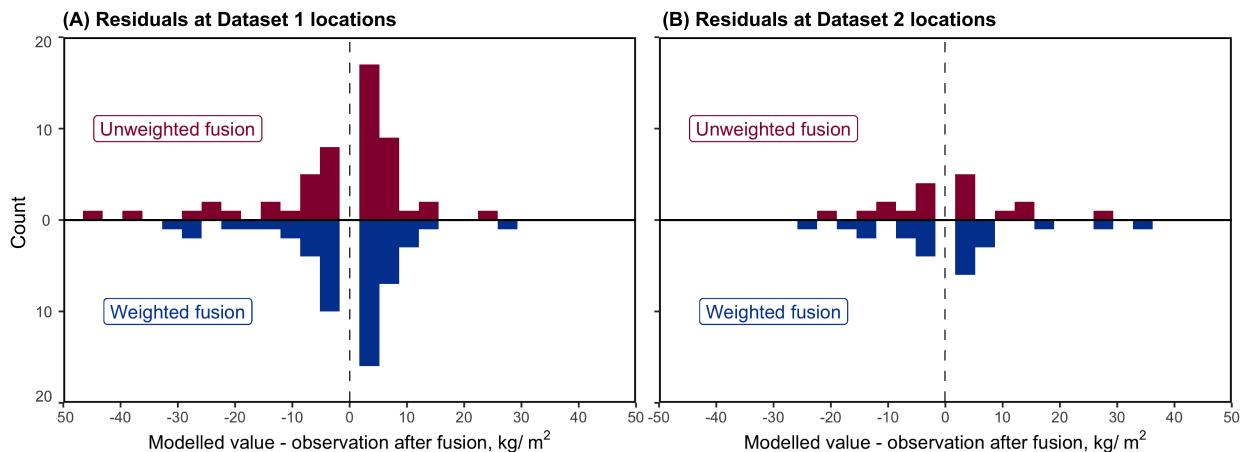


Figure 8: Histograms of residuals at locations of Dataset 1 and 2 after implementing the unweighted and weighted fusion approach. The results show the influence of the weighted fusion on the range/variability of the residuals for the two datasets.

In our analysis, we checked that the assumption of isotropy was not violated (Section A.7). We also confirmed the strength of the spatial correlation, which can tell us if kriging would be suitable as an interpolator and if the form of residual is appropriate. Note that extensions towards non-stationary and anisotropic distributions can be made but are out of scope for this paper.

6. Conclusions

We demonstrate several methodological improvements in the inversion and forward estimation modelling processes when using limited, uncertain and spatial data. We test these using a case study of the eruption of the Kelud volcano in 2014. Using data on tephra thickness collected from two field surveys, we conduct a model inversion to estimate the eruption source parameters and use these in a forward model to estimate the spatial distribution of tephra in locations where no data was collected. We make no specific conclusions about the eruption itself or the true distribution of tephra, but instead, use the test case to demonstrate important considerations and approaches to treating data in inversion-forward estimation problems. Our methodological contributions come in three components: (1) the selection of appropriate cost functions accounting for their behaviour and their associated implied distribution of residuals, (2) the treatment of differential uncertainty when combining multiple data sets, and (3) the leveraging of both forward model and data when estimating the spatial distribution of output.

In the inversion setting, this study places attention on the choice of the cost function to define the best fit between modelled output and observations. This modelling choice receives little attention but has a significant impact on the results as seen in Figure 5. We demonstrated how the characteristics of cost functions and their inherent assumptions on the distribution of model residuals can be used to inform the selection of appropriate cost functions. We also demonstrate the use of data-weighting in the inversion process to account for the differential uncertainty in data. Varying levels of uncertainties associated with different measurements are taken into account by weighting each observation's contribution to the cost function accordingly.

In the forward estimation, we propose to combine the forward modelled output with measured data using spatial statistics techniques. The output is an updated map of tephra fallout that accounts for both the model and the observations. Both the uncertainty in tephra measurements and their spatial correlation are accounted for in the fusion approach. The importance placed in the model and data can be balanced by the choice of the spatial model and its parameters.

One of the strengths of the weighting approach in both inversion and forward model settings is that, where multiple datasets are available, the approach enables making use of all data, even if some are highly uncertain, while accounting for this uncertainty. With the increasing availability of high volume, low reliability information (e.g. crowd-sourced data from social media), these can nonetheless be used to improve the modelling.

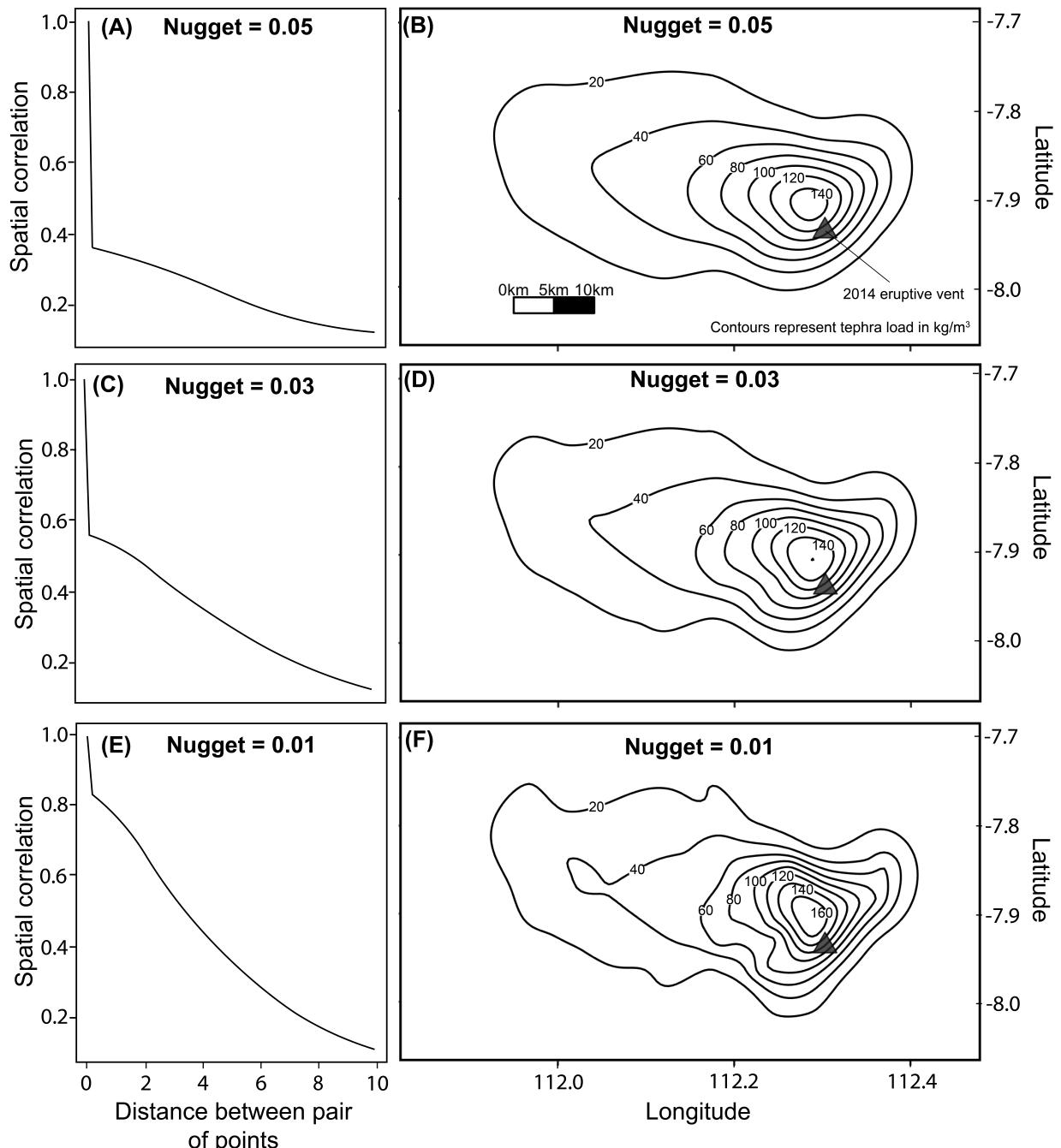


Figure 9: Tephra deposit maps generated from a weighted fusion of the model in Figure 7A and data. Each row of figures correspond to different nuggets used for the spatial model. Figures A, C and E show the spatial correlation structure for different nuggets. Next to the correlation plots are the corresponding tephra deposit maps (Figures B, D and F).

This work serves applications where the inversion and forward prediction workflows are conducted sequentially, and applications where they are done separately. While the test case is focused on *Tephra2*, the methods are applicable to other models that reconstruct the process of tephra fallout such as ASHFALL and FALL3D (Hurst and Turner, 1999; Folch et al., 2009). Beyond applications related to tephra fallout, the guidance could be relevant to those conducting any inversion or calibration modelling using spatial data in earth, environment, and hazards analysis.

Code and data availability

To enable the use of different cost functions and weighted cost functions, we have modified the Tephra2 source code (Version 2.0, Updated 01-27-2018) downloaded from <https://github.com/geoscience-community-codes/tephra2> at the time of writing.

The modified Tephra2 scripts, processed data, and the code developed in R 4.0.2 (R Core Team, 2020a) for the modelling and analysis presented in this paper are available in <https://github.com/ntu-dasl-sg/tephra-spatial-public>.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

MR, MN and DL designed the research. MR wrote the manuscript. SB provided guidance in conducting tephra inversion modelling. All authors discussed the results and commented on the manuscript.

Acknowledgments

This project is supported by the National Research Foundation, Prime Minister's Office, Singapore under the NRF-NRFF2018-06 award, the Earth Observatory of Singapore, the National Research Foundation of Singapore, and the Singapore Ministry of Education under the Research Centers of Excellence initiative. MR is supported by a PhD scholarship from the Earth Observatory of Singapore. We are grateful to Edwin Tan for his support in the use of the High-Performance Computing Cluster, Komodo, in the Earth Observatory of Singapore. We thank George Williams for his support and insights related to the tephra load datasets.

A. Appendix

A.1. Q-Q plots for MAE and MAPE

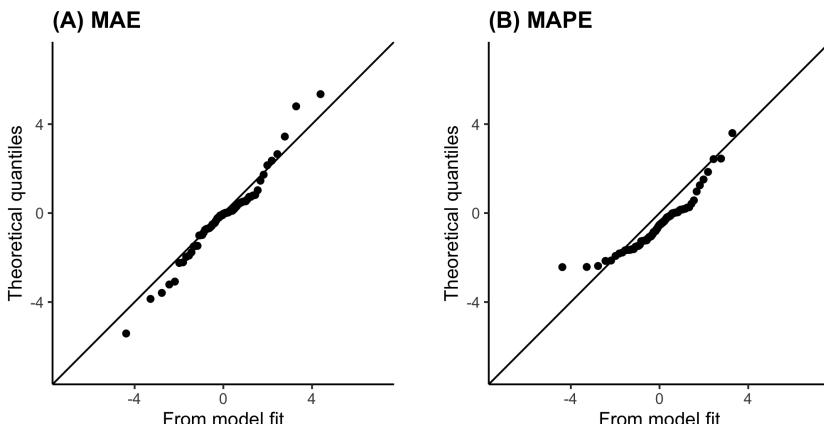


Figure A.1: Quantile-quantile (Q-Q) plots illustrating goodness-of-fit for the different models fitted using different cost functions: mean absolute error (MAE) and mean absolute percentage error (MAPE). The fit is evaluated by how well the empirical quantiles from the transformed model residuals line up with the theoretical quantiles from the assumed distributions along the diagonal line.

A.2. MSLE vs. MAPE

A widely-used order-dependent cost function is the mean absolute percentage error (MAPE), which is obtained by aggregating the relative errors using the mean and then converting to a percentage:

$$\frac{100}{n} \sum_{i=1}^n \left| \frac{\varepsilon_i}{x_i} \right| \quad (\text{A.1})$$

MAPE is used in different areas of research such as atmospheric science and space science (e.g., Grillakis et al. (2013); Zheng and Rosenfeld (2015); Zhelavskaya et al. (2016)). Despite being easy to interpret due to its usage of the percentage, MAPE has been proven to have problems when used in model calibration:

1. MAPE is asymmetric with respect to overestimation and underestimation (Hyndman & Koehler, 2006; Makridakis, 1993; Tofallis, 2015).
2. MAPE is constrained to be positive, so its distribution is generally positively skewed (Hyndman & Koehler, 2006; Swanson et al., 2000).
3. MAPE becomes undefined when the true value is zero (Hyndman & Koehler, 2006).
4. MAPE is not resistant to outliers (Swanson et al., 2000; Tofallis, 2015).

To elaborate on the first point, a prediction of 100 where the observed value is 50 gives a different magnitude of error (100%) than a prediction of 50 where the observed value is 100 (50%). Underprediction is therefore less heavily penalized than overprediction, even if the order of the error is the same. The third point also means that MAPE is not an appropriate metric where the quantity being predicted is likely to be zero (e.g., Tofallis (2015)).

A.3. Simple kriging interpolation

The procedure for simple kriging interpolation is as follows:

1. For each pair of residuals at the sampled sites, calculate the variogram value γ as half of the mean-squared difference between their values.
2. Define a model of spatial variation (i.e. variogram model) that best relates the variogram values from Step 1, and the corresponding separation distances (h) for each pair of locations. The variogram model's parameters may be user-defined or estimated using a maximum likelihood estimation approach. In defining the variogram model, we adopt the Matérn function, and use the geostats and geoR packages in R programming environment for fitting (Diggle and Jr, 2007; Ribeiro Jr et al., 2020; Vermeesch, 2022; R Core Team, 2020b).
3. Estimate the elements of the covariance matrix, Σ , which describes the spatial relationships among the values at the observed sites and the prediction location. The covariance matrix has a size of $N \times N$, and its $i - j$ th element is defined as

$$c(s_i, s_j) = \sigma^2 \gamma(h) + \tau^2, \quad (\text{A.2})$$

where σ^2 is the amount of spatially correlated variation, $\gamma(h)$ is the variogram model from Step 2, τ^2 is the variogram value at zero separation distance, and h is the distance between the two observation sites s_i and s_j . The parameters σ^2 and τ^2 are derived from the variogram model. The variable τ^2 is more commonly known as the *nugget* in the geostatistics literature, and it represents the variability in the interpolated at very short separation distances.

4. Calculate kriging weights λ_i as:

$$\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix} = \begin{pmatrix} c(s_1, s_1) & \cdots & c(s_1, s_N) \\ \vdots & \ddots & \vdots \\ c(s_N, s_1) & \cdots & c(s_N, s_N) \end{pmatrix}^{-1} \begin{pmatrix} c(s_1, s_0) \\ \vdots \\ c(s_N, s_0) \end{pmatrix} \quad (\text{A.3})$$

5. Calculate the residual for all sites in the study area. The residual at any prediction location s_o can be written as:

$$\hat{Z}(s_o) = \sum_{i=1}^N \lambda_i Z(s_i) \quad (\text{A.4})$$

where $Z(s_i)$ is the residual at an observed location s_i calculated using the equations for Z_i in Table 2, λ_i is the weight applied to the value at the observed location, and N is the number of observed sites.

6. The associated kriging error can be calculated as:

$$Var(\hat{Z}(s_o) - Z(s_o)) = c(s_0, s_0) - \begin{pmatrix} c(s_1, s_0) \\ \vdots \\ c(s_N, s_0) \end{pmatrix}' \begin{pmatrix} c(s_1, s_1) & \cdots & c(s_1, s_N) \\ \vdots & \ddots & \vdots \\ c(s_N, s_1) & \cdots & c(s_N, s_N) \end{pmatrix}^{-1} \begin{pmatrix} c(s_1, s_0) \\ \vdots \\ c(s_N, s_0) \end{pmatrix} \quad (\text{A.5})$$

A.4. Back transformation

If MSE was used as the cost function in the inversion, the final prediction map can be written as $\hat{Y} = \mu + \hat{\epsilon}$ where μ denotes the forward model prediction. The associated spatial prediction uncertainty $Var[\hat{Y}(\mathbf{x})]$ is equivalent to $Var[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]$.

For the other cost functions related to Gaussian distributions, namely the chi-square error and MSLE, we can transform the residuals as outlined in Table 3 before applying simple kriging. We would then need to back-transform the predicted residuals and their prediction errors to obtain the final prediction maps.

For chi-square loss:

$$\hat{Y} = \mu + \sqrt{\mu}\hat{\epsilon} \text{ and } Var[\hat{Y}] = \mu Var[\hat{\epsilon}]. \quad (\text{A.6})$$

For MSLE:

$$\hat{Y} = \exp\left(\frac{\log_{10}(\mu + 1) - \hat{\epsilon}}{\log_{10}e} + \frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{2(\log_{10}e)^2}\right) - 1 \quad (\text{A.7})$$

$$\text{and } Var[\hat{Y}] = \left[\exp\left(\frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{(\log_{10}e)^2}\right) - 1 \right] \exp\left(2\frac{\log_{10}(\mu + 1) - \hat{\epsilon}}{\log_{10}e} + \frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{(\log_{10}e)^2}\right) \quad (\text{A.8})$$

A derivation of the prediction formulae for MSLE is given in Section A.5.

A.5. Derivation of the MSLE prediction formulae

When we use mean squared logarithmic error (MSLE) as a loss function to calibrate our forward model, we are implicitly assuming a Gaussian distribution on its transformed residual $Z_i = \log_{10}\left(\frac{\text{Predicted}+1}{\text{Actual}+1}\right)$ where *Predicted* = μ is the forward model prediction and *Actual* is the observed tephra load (see Table 2 in the Main Paper). Hence, to model spatial deviations from the forward model, we conduct simple kriging or the method in Worden et al. (2018) on the transformed residuals.

Let $\hat{\epsilon} = \mathbb{E}[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]$ denote the kriging prediction and $[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]$ the prediction variance. Then:

$$\log_{10}(\text{Actual} + 1) \sim N(\log_{10}(\text{Predicted} + 1) - \hat{\epsilon}, [\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]), \quad (\text{A.9})$$

$$\text{since } \log_{10}(\text{Actual} + 1) = \log_{10}(\text{Predicted} + 1) - \log_{10}\left(\frac{\text{Predicted}+1}{\text{Actual}+1}\right).$$

Change bases from 10 to e , we have:

$$\log_e(\text{Actual} + 1) = \frac{\log_{10}(\text{Actual} + 1)}{\log_{10}(e)} \sim N\left(\frac{\log_{10}(\text{Predicted} + 1) - \hat{\epsilon}}{\log_{10}(e)}, \frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{(\log_{10}(e))^2}\right), \quad (\text{A.10})$$

This means that $\text{Actual} + 1 \sim \text{logNormal}\left(\frac{\log_{10}(\text{Predicted}+1)-\hat{\epsilon}}{\log_{10}(e)}, \frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{(\log_{10}(e))^2}\right)$. Using the formulae for the mean and the variance of a logNormal random variable, and writing *Predicted* as μ , the estimated tephra load and its prediction variance are:

$$\hat{Y} = \exp\left(\frac{\log_{10}(\mu + 1) - \hat{\epsilon}}{\log_{10}e} + \frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{2(\log_{10}e)^2}\right) - 1 \quad (\text{A.11})$$

$$\text{and } [\hat{Y}] = \left[\exp\left(\frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{(\log_{10} e)^2} \right) - 1 \right] \exp\left(2 \frac{\log_{10}(\mu + 1) - \hat{\epsilon}}{\log_{10} e} + \frac{[\epsilon(\mathbf{x})|\epsilon(\mathbf{x}_1), \dots, \epsilon(\mathbf{x}_N)]}{(\log_{10} e)^2} \right). \quad (\text{A.12})$$

A.6. LOOCV procedure

LOOCV computes the test performance for all reference data points. The following steps are conducted for both the simple kriging in Section 4.4.1, and its extension in Section 4.4.2 which accounts for varying data uncertainty:

1. We make use of the transformed residuals at the observation sites calculated in Section 4.4.1 Step 2a. For brevity in the steps that follow, we refer to the transformed residual at an observation site as *observation point*.
2. We set aside one observation point as a test point, while the other observation points form the training set. Note that we only use observation points from the reference dataset (i.e. Dataset 1) as test set. The idea is to evaluate the performance of the interpolation approach on the most reliable data.
3. Using only the training points, kriging interpolation is conducted for the test point.
4. This interpolated value of transformed residual is combined to the optimised model estimate for the test location following Section 4.4.1's Steps 3a and 3b. The result is an updated value of tephra load at test site.
5. The above steps are repeated until all points from Dataset 1 are evaluated as a test point. For all iterations, the interpolated load estimates are compared against the observed tephra load values based on a performance metric (e.g. root mean square error, chi-square, and mean square log error). The result is the LOOCV error associated with the kriging interpolation approach used in Step 3.

A.7. Checks for isotropy and residual form

We investigate the strength of the spatial correlation when the residuals are transformed in different forms, with the aim of finding which residual form is best for the kriging methods. We hypothesize that the log-residual is more suitable for use in kriging, since in Section 5.1, we identified that MSLE is the most suitable cost function for the case study.

In Figure A.2, we show the empirical and fitted variograms for the different forms of residuals. Annotated in the figures are the variograms' *range*, which vary across the variogram models shown. For instance, the range up to which log residuals can be interpolated (around 22km) was around three times larger than for the actual residuals (around 7km). This indicates that the log residuals are more appropriate to interpolate the residuals. Because of this evidence, we proceed with using log residuals for the spatial interpolation of the model deviations.

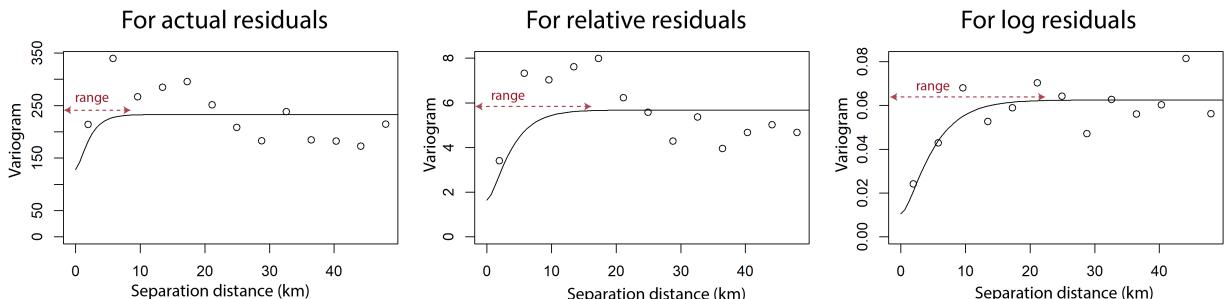


Figure A.2: Empirical and fitted Matérn variogram for actual and log residuals. Note that the widely different range of semivariance values in the y-axis is due to the scale of which the residuals are transformed. The range of the variograms (indicated with a dotted line) indicate the lag distances where data is spatially correlated. Beyond the range, interpolation does not occur.

We also confirmed the validity of assuming an isotropic field when modelling the spatial correlation, for which we use directional variograms. This step serves as a statistical check of the assumptions made in the kriging methodology since exterior factors (i.e. related to the physics of tephra deposition) may affect the spatial structure of the dataset and the model deviations. To generate the directional variograms, the variance for pairs of data were calculated for specific directions within a certain tolerance. Since the correlation structure appears similar for all the directional variograms in Figure A.3, there's no evidence of anisotropy. Hence, the assumption of an isotropic field is acceptable.

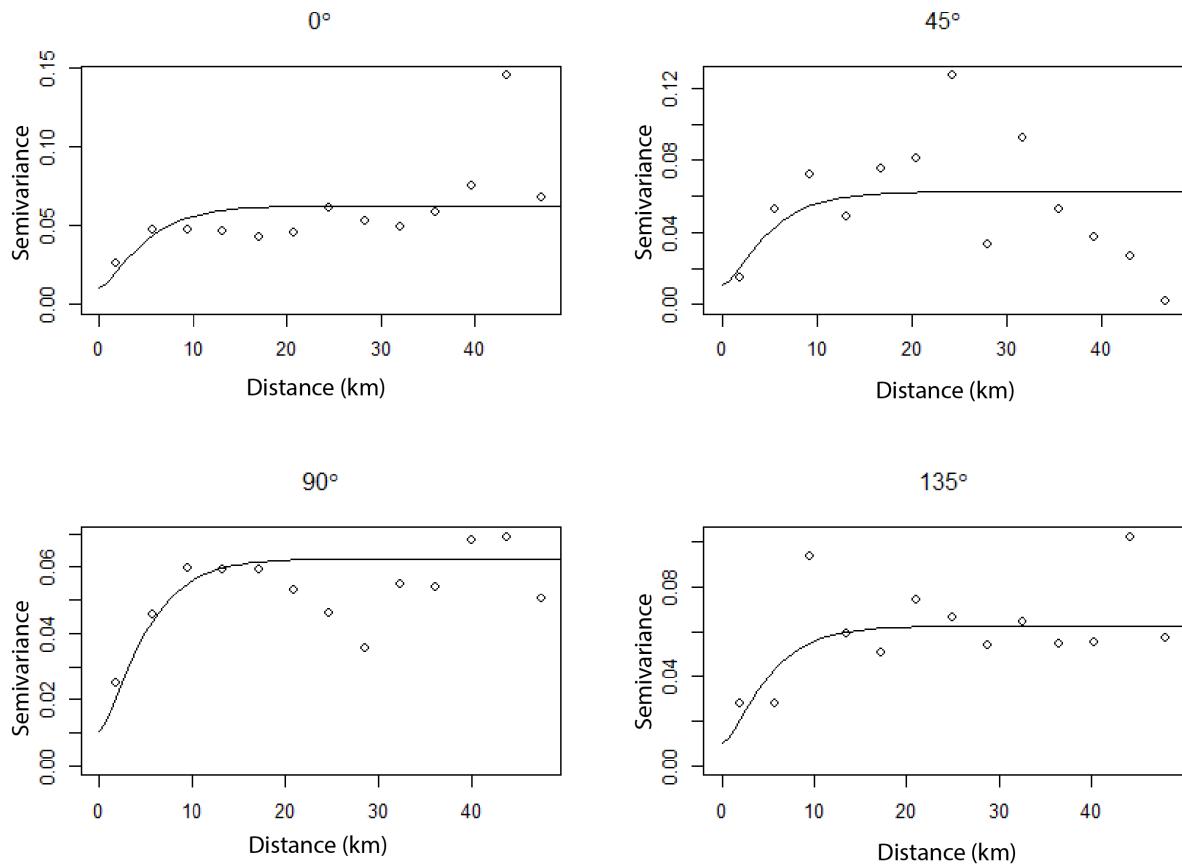


Figure A.3: The fitted omnidirectional Matérn model (shown in solid line) and empirical directional variograms (points) for the directions N0E, N45E, N90E, and N135E. An angular tolerance of 22.5 is set for generating the directional variograms.

A.8. Leave-one-out cross-validation errors from fusion with different variogram models

Table A.1 shows the leave-one-out cross-validation (LOOCV) errors after weighted and unweighted fusion. For all cost functions and variogram models considered except the case of the Gaussian model for the chi-square cost function, the errors are lower for weighted fusion as compared to those from the unweighted fusion. In addition, the LOOCV errors are lower than the pre-kriging training error, indicating that there was no overfitting due to kriging.

A.9. Standard deviation of the tephra load estimates from model fusion

In Figure A.4, we show the maps of standard deviations associated with the tephra load estimates from the weighted and unweighted fusion. Note that the standard deviations are higher for larger estimated loads because the kriging was performed on a relative, logarithmic scale.

References

- Anggorowati, A., Harijoko, A., 2015. Distribusi area, volume, serta karakteristik mineralogi dan geokimia endapan tefra jatuh dari erupsi gunung kelud tahun 2014, in: Seminar Nasional Kebumian KE-8: Academia-Industry Linkage, p. 778–789. URL: <https://repository.ugm.ac.id/135520/>.
- Armienti, P., Macedonio, G., Pareschi, M., 1988. A numerical model for simulation of tephra transport and deposition: Applications to may 18, 1980, mount st. helens eruption. Journal of Geophysical Research: Solid Earth 93, 6463–6476.
- Armstrong, J.S., Fildes, R., 1995. Correspondence on the selection of error measures for comparisons among forecasting methods. Journal of Forecasting 14, 67–71.
- Arnalds, O., Thorarinsson, E.F., Thorsson, J., Waldhauserova, P.D., Agustsdottir, A.M., 2013. An extreme wind erosion event of the fresh Eyjafjallajökull 2010 volcanic ash. Scientific reports 3, 1–7.

Table A.1

Measures of performance of the unweighted and weighted kriging-based fusion methods in terms of LOOCV errors for different variogram models. The errors were calculated using three different metrics with formulas indicated in the table: root mean squared errors, chi-square errors and mean square log error. The training errors before applying the fusion methods are consistently higher than the LOOCV errors, which indicates no overfitting in the kriging method.

Cost function	Performance metric	Variogram model	Pre-kriging training error	LOOCV error after unweighted fusion	LOOCV error after weighted kriging
MSE	Root-mean-square	Matérn		15.470	15.064
		Exponential	16.166	15.307	14.984
		Gaussian		15.924	15.717
Chi-square	Chi-square	Matérn		5.362	5.300
		Exponential	8.051	5.404	5.359
		Gaussian		5.494	5.533
MSLE	Mean square log	Matérn		0.040	0.033
		Exponential	0.0570	0.0395	0.034
		Gaussian		0.040	0.034

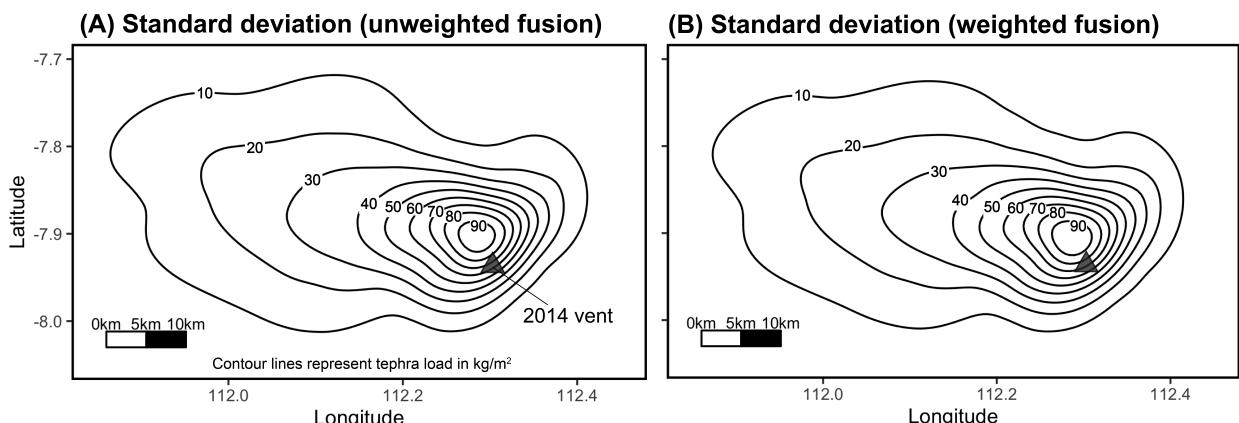


Figure A.4: Map of standard deviations associated with the estimates of tephra load from the fusion of the best-fit modelled output (using MSLE as cost function) shown in Figure 7A. The map in A corresponds to the unweighted fusion estimates in Figure 7B while B is associated with the weighted fusion estimates in Figure 9B.

- Ayris, P.M., Delmelle, P., 2012. The immediate environmental effects of tephra emission. *Bulletin of Volcanology* 74, 1905–1936.
- Bear-Crozier, A., Pouget, S., Bursik, M., Jansons, E., Denman, J., Tupper, A., Rustowicz, R., 2020. Automated detection and measurement of volcanic cloud growth: towards a robust estimate of mass flux, mass loading and eruption duration. *Natural Hazards* 101, 1–38.
- Biass, S., Todde, A., Cioni, R., Pistolesi, M., Geshi, N., Bonadonna, C., 2017. Potential impacts of tephra fallout from a large-scale explosive eruption at Sakurajima volcano, Japan. *Bulletin of Volcanology* 79, 1–24.
- Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Marzouk, Y., Tenorio, L., van Bloemen Waanders, B., Willcox, K., 2011. Large-scale inverse problems and quantification of uncertainty.
- Blong, R., Enright, N., Grasso, P., 2017. Preservation of thin tephra. *Journal of Applied Volcanology* 6, 10.
- Bonadonna, C., Biass, S., Costa, A., 2015. Physical characterization of explosive volcanic eruptions based on tephra deposits: propagation of uncertainties and sensitivity analysis. *Journal of Volcanology and Geothermal Research* 296, 80–100.
- Bonadonna, C., Connor, C.B., Houghton, B., Connor, L., Byrne, M., Laing, A., Hincks, T., 2005. Probabilistic modeling of tephra dispersal: Hazard assessment of a multiphase rhyolitic eruption at Tarawera, New Zealand. *Journal of Geophysical Research: Solid Earth* 110.
- Bonadonna, C., Costa, A., 2013. Modeling of tephra sedimentation from volcanic plumes. *Modeling volcanic processes: The physics and mathematics of volcanism*, 173–202.
- Cabaneros, S.M., Hughes, B., 2022. Methods used for handling and quantifying model uncertainty of artificial neural network models for air pollution forecasting. *Environmental Modelling & Software*, 105529.
- Carey, S., Sparks, R., 1986. Quantitative models of the fallout and dispersal of tephra from volcanic eruption columns. *Bulletin of Volcanology* 48, 109–125.

- Caudron, C., Taisne, B., Garcés, M., Alexis, L.P., Mialle, P., 2015. On the use of remote infrasound and seismic stations to constrain the eruptive sequence and intensity for the 2014 kelud eruption. *Geophysical Research Letters* 42, 6614–6621. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL064885>, doi:<https://doi.org/10.1002/2015GL064885>, arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015GL064885>.
- Chen, C., Twycross, J., Garibaldi, J.M., 2017. A new accuracy measure based on bounded relative error for time series forecasting. *PloS one* 12, e0174202.
- Connor, C.B., Connor, L.J., Bonadonna, C., Luhr, J., Savov, I., Navarro-Ochoa, C., 2019. Modelling tephra thickness and particle size distribution of the 1913 eruption of volcán de colima, mexico, in: *Volcán de Colima*. Springer, pp. 81–110.
- Connor, L., Connor, C., Saballos, A., 2011. Tephra2 users manual. University of South Florida, Tampa, FL, accessed Sept 24, 2017.
- Connor, L.J., Connor, C.B., 2006. Inversion is the key to dispersion: understanding eruption dynamics by inverting tephra fallout, in: *Statistics in Volcanology*. Geological Society of London.
- Constantinescu, R., Hopule-Gligor, A., Connor, C.B., Bonadonna, C., Connor, L.J., Lindsay, J.M., Charbonnier, S., Volentik, A.C., 2021. The radius of the umbrella cloud helps characterize large explosive volcanic eruptions. *Communications Earth & Environment* 2, 1–8.
- Costa, A., Folch, A., Macedonio, G., Giaccio, B., Isaia, R., Smith, V., 2012. Quantifying volcanic ash dispersal and impact of the campanian ignimbrite super-eruption. *Geophysical Research Letters* 39.
- Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d.P., et al., 2011. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* 137, 553–597.
- Diggle, P.J., Jr, P.J.R., 2007. *Model Based Geostatistics*. Springer, New York.
- Dominguez, L., Bonadonna, C., Forte, P., Jarvis, P.A., Cioni, R., Mingari, L., Bran, D., Panebianco, J.E., 2020. Aeolian remobilisation of the 2011-Cordón Caulle Tephra-Fallout Deposit: example of an important process in the life cycle of Volcanic Ash. *Frontiers in Earth Science* 7, 343.
- Engle, R.F., 1993. On the limitations of comparing mean square forecast errors: Comment. *Journal of Forecasting* 12, 642–644.
- Engwell, S., Sparks, R., Aspinall, W., 2013. Quantifying uncertainties in the measurement of tephra fall thickness. *Journal of Applied Volcanology* 2, 1–12.
- Folch, A., Costa, A., Macedonio, G., 2009. Fall3d: A computational model for transport and deposition of volcanic ash. *Computers & Geosciences* 35, 1334–1342.
- Fontijn, K., Williamson, D., Mbede, E., Ernst, G.G., 2012. The rungwe volcanic province, tanzania—a volcanological review. *Journal of African Earth Sciences* 63, 12–31.
- Georgoudas, I.G., Sirakoulis, G.C., Scordilis, E.M., Andreadis, I., 2007. A cellular automaton simulation tool for modelling seismicity in the region of xanthi. *Environmental Modelling & Software* 22, 1455–1464.
- Global Volcanism Program, 2014. Report on kelut (indonesia). *Bulletin of the Global Volcanism Network* 39. doi:<https://doi.org/10.5479/si.GVP.BGVN201402-263280>.
- Gómez-Romero, M., Lindig-Cisneros, R., Galindo-Vallejo, S., 2006. Effect of tephra depth on vegetation development in areas affected by volcanism. *Plant Ecology* 183, 207–213.
- Goode, L.R., Handley, H.K., Cronin, S.J., Abdurrahman, M., 2019. Insights into eruption dynamics from the 2014 pyroclastic deposits of kelut volcano, java, indonesia, and implications for future hazards. *Journal of Volcanology and Geothermal Research* 382, 6–23.
- Grillakis, M.G., Koutroulis, A.G., Tsanis, I.K., 2013. Multisegment statistical bias correction of daily gcm precipitation output. *Journal of Geophysical Research: Atmospheres* 118, 3150–3162.
- Hargie, K.A., Van Eaton, A.R., Mastin, L.G., Holzworth, R.H., Ewert, J.W., Pavolonis, M., 2019. Globally detected volcanic lightning and umbrella dynamics during the 2014 eruption of kelud, indonesia. *Journal of Volcanology and Geothermal Research* 382, 81–91. URL: <https://www.sciencedirect.com/science/article/pii/S0377027318304591>, doi:<https://doi.org/10.1016/j.jvolgeores.2018.10.016>. lessons learned from the recent eruptions of Sinabung and Kelud Volcanoes, Indonesia.
- Hayes, J.L., Calderón B, R., Deligne, N.I., Jenkins, S.F., Leonard, G.S., McSpadden, A.M., Williams, G.T., Wilson, T.M., 2019. Timber-framed building damage from tephra fall and lahar: 2015 Calbuco eruption, Chile. *Journal of Volcanology and Geothermal Research* 374, 142–159. URL: <https://www.sciencedirect.com/science/article/pii/S0377027318303718>, doi:<https://doi.org/10.1016/j.jvolgeores.2019.02.017>.
- Hayes, S.K., Montgomery, D.R., Newhall, C.G., 2002. Fluvial sediment transport and deposition following the 1991 eruption of mount pinatubo. *Geomorphology* 45, 211–224.
- Hodson, T.O., 2022. Root mean square error (rmse) or mean absolute error (mae): when to use them or not. *Geoscientific Model Development Discussions* , 1–10.
- Hollós, R., Fodor, N., Merganičová, K., Hidy, D., Árendás, T., Grünwald, T., Barcza, Z., 2022. Conditional interval reduction method: A possible new direction for the optimization of process based models. *Environmental Modelling & Software* 158, 105556.
- Hurst, A., Turner, R., 1999. Performance of the program ashfall for forecasting ashfall during the 1995 and 1996 eruptions of ruapehu volcano. *New Zealand Journal of Geology and Geophysics* 42, 615–622.
- IFRC, 2014. Emergency plan of action (epoa) indonesia volcanic eruption-mt kelud, in: Situation Report by the International Federation of Red Cross And Red Crescent Societies. URL: <https://reliefweb.int/report/indonesia/indonesia-volcanic-eruption-mt-kelud-emergency-plan-action-epoa-operation-n>.
- Isaac, T., Petra, N., Stadler, G., Ghantas, O., 2015. Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the antarctic ice sheet. *Journal of Computational Physics* 296, 348–368. URL: <https://www.sciencedirect.com/science/article/pii/S0021999115003046>, doi:<https://doi.org/10.1016/j.jcp.2015.04.047>.

- Kim, S., Kim, H., 2016. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting* 32, 669–679.
- Kristiansen, N.I., Prata, A., Stohl, A., Carn, S.A., 2015. Stratospheric volcanic ash emissions from the 13 february 2014 kelut eruption. *Geophysical Research Letters* 42, 588–596.
- Le Pennec, J.L., Ruiz, G.A., Ramón, P., Palacios, E., Mothes, P., Yépez, H., 2012. Impact of tephra falls on Andean communities: The influences of eruption size and weather conditions during the 1999–2001 activity of Tungurahua volcano, Ecuador. *Journal of Volcanology and Geothermal Research* 217, 91–103.
- Li, Z., Duque, F.Q., Grout, T., Bates, B., Demir, I., 2022. Comparative analysis of performance and mechanisms of flood inundation map generation using height above nearest drainage .
- Madankan, R., Pouget, S., Singla, P., Bursik, M., Dehn, J., Jones, M., Patra, A., Pavoloni, M., Pitman, E.B., Singh, T., et al., 2014. Computation of probabilistic hazard maps and source parameter estimation for volcanic ash transport and dispersion. *Journal of Computational Physics* 271, 39–59.
- Maeno, F., Nakada, S., Yoshimoto, M., Shimano, T., Hokanishi, N., Zaennudin, A., Iguchi, M., 2019. A sequence of a plinian eruption preceded by dome destruction at kelud volcano, indonesia, on february 13, 2014, revealed from tephra fallout and pyroclastic density current deposits. *Journal of Volcanology and Geothermal Research* 382, 24 – 41. URL: <http://www.sciencedirect.com/science/article/pii/S0377027317301385>, doi:<https://doi.org/10.1016/j.jvolgeores.2017.03.002>. lessons learned from the recent eruptions of Sinabung and Kelud Volcanoes, Indonesia.
- Magill, C., Wilson, T., Okada, T., 2013. Observations of tephra fall impacts from the 2011 Shinmoedake eruption, Japan. *Earth, Planets and Space* 65, 677–698.
- Makridakis, S., 1993. Accuracy measures: theoretical and practical concerns. *International journal of forecasting* 9, 527–529.
- Mannen, K., 2014. Particle segregation of an eruption plume as revealed by a comprehensive analysis of tephra dispersal: theory and application. *Journal of volcanology and geothermal research* 284, 61–78.
- Mastin, L.G., Guffanti, M., Servranckx, R., Webley, P., Barsotti, S., Dean, K., Durant, A., Ewert, J.W., Neri, A., Rose, W.I., et al., 2009. A multidisciplinary effort to assign realistic source parameters to models of volcanic ash-cloud transport and dispersion during eruptions. *Journal of Volcanology and Geothermal Research* 186, 10–21.
- Morley, S.K., Brito, T.V., Welling, D.T., 2018. Measures of model performance based on the log accuracy ratio. *Space Weather* 16, 69–88.
- Newhall, C.G., Self, S., 1982. The volcanic explosivity index (vei) an estimate of explosive magnitude for historical volcanism. *Journal of Geophysical Research: Oceans* 87, 1231–1238.
- Oden, T., Moser, R., Ghattas, O., 2010. Computer predictions with quantified uncertainty, part i. *SIAM News* 43, 1–3.
- Oishi, M., Nishiki, K., Geshi, N., Furukawa, R., Ishizuka, Y., Oikawa, T., Yamamoto, T., Nanayama, F., Tanaka, A., Hirota, A., et al., 2018. Distribution and mass of tephra-fall deposits from volcanic eruptions of Sakurajima Volcano based on posteruption surveys. *Bulletin of Volcanology* 80, 1–16.
- Pieri, D.C., Baloga, S.M., 1986. Eruption rate, area, and length relationships for some hawaiian lava flows. *Journal of volcanology and geothermal research* 30, 29–45.
- Pouget, S., Bursik, M., Webley, P., Dehn, J., Pavoloni, M., 2013. Estimation of eruption source parameters from umbrella cloud or downwind plume growth rate. *Journal of volcanology and geothermal research* 258, 100–112.
- Pyle, D., 2016. Chapter 1 - field observations of tephra fallout deposits, in: Mackie, S., Cashman, K., Ricketts, H., Rust, A., Watson, M. (Eds.), *Volcanic Ash*. Elsevier, pp. 25–37. URL: <https://www.sciencedirect.com/science/article/pii/B9780081004050000045>, doi:<https://doi.org/10.1016/B978-0-08-100405-0-00004-5>.
- Pyle, D.M., 1989. The thickness, volume and grainsize of tephra fall deposits. *Bulletin of Volcanology* 51, 1–15.
- R Core Team, 2020a. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- R Core Team, 2020b. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Ramachandran, K.M., Tsokos, C.P., 2015. Chapter 7 - goodness-of-fit tests applications, in: Ramachandran, K.M., Tsokos, C.P. (Eds.), *Mathematical Statistics with Applications in R*. 2nd ed.. Academic Press, Boston, pp. 371–408. URL: <https://www.sciencedirect.com/science/article/pii/B9780124171138000072>, doi:<https://doi.org/10.1016/B978-0-12-417113-8.00007-2>.
- Ribeiro Jr, P.J., Diggle, P.J., Schlather, M., Bivand, R., Ripley, B., 2020. geoR: Analysis of Geostatistical Data. URL: <https://CRAN.R-project.org/package=geoR>. r package version 1.8-1.
- Scarpati, C., Cole, P., Perrotta, A., 1993. The neapolitan yellow tuff—a large volume multiphase eruption from campi flegrei, southern italy. *Bulletin of Volcanology* 55, 343–356.
- Scollo, S., Folch, A., Costa, A., 2008. A parametric and comparative study of different tephra fallout models. *Journal of Volcanology and Geothermal Research* 176, 199–211.
- Shao, J., 1993. Linear model selection by cross-validation. *Journal of the American statistical Association* 88, 486–494.
- Spence, R.J.S., Kelman, I., Baxter, P.J., Zuccaro, G., Petrazzoli, S., 2005. Residential building and occupant vulnerability to tephra fall. *Natural Hazards and Earth System Sciences* 5, 477–494. URL: <https://nhess.copernicus.org/articles/5/477/2005/>, doi:10.5194/nhess-5-477-2005.
- Stephens, M., 1986. Tests based on edf statistics, chapter 4, in: D'Agostino, R., Stephens, M. (Eds.), *Goodness-of-Fit Techniques*. 1st ed.. Taylor & Francis Group, Boca Raton, pp. 97–104.
- Stohl, A., Prata, A., Eckhardt, S., Clarisse, L., Durant, A., Henne, S., Kristiansen, N.I., Minikin, A., Schumann, U., Seibert, P., et al., 2011. Determination of time-and height-resolved volcanic ash emissions and their use for quantitative ash dispersion modeling: the 2010 eyjafjallajökull eruption. *Atmospheric Chemistry and Physics* 11, 4333–4351.
- Suzuki, T., et al., 1983. A theoretical model for dispersion of tephra. *Arc volcanism: physics and tectonics* 95, 113.

Inversion-forward estimation of tephra fallout

- Suzuki, Y., Iguchi, M., 2019. Determination of the mass eruption rate for the 2014 mount kelud eruption using three-dimensional numerical simulations of volcanic plumes. *Journal of Volcanology and Geothermal Research* 382, 42–49. URL: <https://www.sciencedirect.com/science/article/pii/S0377027317303657>, doi:<https://doi.org/10.1016/j.jvolgeores.2017.06.011>. lessons learned from the recent eruptions of Sinabung and Kelud Volcanoes, Indonesia.
- Tanaka, H.L., Iguchi, M., Nakada, S., 2016. Numerical simulations of volcanic ash plume dispersal from Kelud volcano in Indonesia on February 13, 2014. *Journal of Disaster Research* 11, 31–42.
- Thompson, M.A., Lindsay, J.M., Wilson, T.M., Biass, S., Sandri, L., 2017. Quantifying risk to agriculture from volcanic ashfall: a case study from the Bay of Plenty, New Zealand. *Natural Hazards* 86, 31–56.
- Thornes, J.E., Stephenson, D.B., 2001. How to judge the quality and value of weather forecast products. *Meteorological Applications* 8, 307–314.
- Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society* 66, 1352–1362.
- Uhlenbrook, S., Roser, S., Tilch, N., 2004. Hydrological process representation at the meso-scale: the potential of a distributed, conceptual catchment model. *Journal of Hydrology* 291, 278–296.
- Vermeesch, P., 2022. geostats: An Introduction to Statistics for Geoscientists. URL: <https://CRAN.R-project.org/package=geostats.r> package version 1.5.
- Volentik, A.C., Bonadonna, C., Connor, C.B., Connor, L.J., Rosi, M., 2010. Modeling tephra dispersal in absence of wind: Insights from the climactic phase of the 2450 BP Plinian eruption of Pululagua volcano (Ecuador). *Journal of Volcanology and Geothermal Research* 193, 117–136.
- Wackernagel, H., 2003. *Linear Regression and Simple Kriging*. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 15–26. URL: https://doi.org/10.1007/978-3-662-05294-5_3, doi:[10.1007/978-3-662-05294-5_3](https://doi.org/10.1007/978-3-662-05294-5_3).
- Walker, G., Croasdale, R., 1971. Characteristics of some basaltic pyroclastics. *Bulletin volcanologique* 35, 303–317.
- Walther, B.A., Moore, J.L., 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* 28, 815–829.
- Wang, T., Shen, Y., Xie, X., Chai, J., 2022. Ground-motion simulation using stochastic finite-fault method combined with a parameter calibration process based on historical seismic data. *Natural Hazards*, 1–20.
- Wardman, J., Wilson, T., Bodger, P., Cole, J., Stewart, C., 2012. Potential impacts from tephra fall to electric power systems: a review and mitigation strategies. *Bulletin of volcanology* 74, 2221–2241.
- White, J., Connor, C.B., Connor, L., Hasenaka, T., 2017. Efficient inversion and uncertainty quantification of a tephra fallout model. *Journal of Geophysical Research: Solid Earth* 122, 281–294.
- Willcox, K.E., Ghosh, O., Heimbach, P., 2021. The imperative of physics-based modeling and inverse theory in computational science. *Nature Computational Science* 1, 166–168.
- Williams, G.T., Jenkins, S.F., Biass, S., Wibowo, H.E., Harijoko, A., 2020. Remotely assessing tephra fall building damage and vulnerability: Kelud Volcano, Indonesia. *Journal of Applied Volcanology* 9, 1–18.
- Wilson, T., Cole, J., Stewart, C., Cronin, S., Johnston, D., 2011. Ash storms: impacts of wind-remobilised volcanic ash on rural communities and agriculture following the 1991 Hudson eruption, southern Patagonia, Chile. *Bulletin of Volcanology* 73, 223–239.
- Worden, C., Wald, D., Allen, T., Lin, K., Garcia, D., Cua, G., 2010. A revised ground-motion and intensity interpolation scheme for shakemap. *Bulletin of the Seismological Society of America* 100, 3083–3096.
- Worden, C.B., Thompson, E.M., Baker, J.W., Bradley, B.A., Luco, N., Wald, D.J., 2018. Spatial and spectral interpolation of ground-motion intensity measure observations. *Bulletin of the Seismological Society of America* 108, 866–875.
- Yang, Q., Pitman, E.B., Bursik, M., Jenkins, S.F., 2021. Tephra deposit inversion by coupling tephra2 with the metropolis-hastings algorithm: algorithm introduction and demonstration with synthetic datasets. *Journal of Applied Volcanology* 10, 1–24.
- Zhelavskaya, I., Spasojevic, M., Shprits, Y., Kurth, W., 2016. Automated determination of electron density from electric field measurements on the van allen probes spacecraft. *Journal of Geophysical Research: Space Physics* 121, 4611–4625.
- Zheng, Y., Rosenfeld, D., 2015. Linear relation between convective cloud base height and updrafts and application to satellite retrievals. *Geophysical Research Letters* 42, 6485–6491.