Bulletin of Volcanology (2023) 85:68
https://doi.org/10.1007/s00445-023-01664-x

RESEARCH ARTICLE

Quantifying uncertainty in probabilistic volcanic ash hazard forecasts, with an application to weather pattern based wind field sampling

Jeremy Phillips[1] · Shannon Williams[2] · Anthony Lee[2] · Susanna Jenkins[3]

Received: 12 January 2023 / Accepted: 11 August 2023 / Published online: 31 October 2023
© The Author(s) 2023

Abstract

Probabilistic forecasting of volcanic ash dispersion involves simulating an ensemble of realistic event scenarios to estimate the probability of a particular hazard threshold being exceeded. Although the number of samples that make up the ensemble, how they are chosen, and the desired threshold all set the uncertainty of (or confidence in) the estimated exceedance probability, current practice does not quantify and communicate the uncertainty in ensemble predictions. In this study, we use standard statistical methods to estimate the variance in probabilistic ensembles and use this measure of uncertainty to assess different sampling strategies for the wind field, using the example of volcanic ash transport from a representative explosive eruption in Iceland. For stochastic (random) sampling of the wind field, we show how the variance is reduced with increasing ensemble size and how the variance depends on the desired hazard threshold and the proximity of a target site to the volcanic source. We demonstrate how estimated variances can be used to compare different ensemble designs, by comparing stochastic forecasts with forecasts obtained from a stratified sampling approach using a set of 29 Northern European weather regimes, known as Grosswetterlagen (GWL). Sampling wind fields from within the GWL regimes reduces the number of samples needed to achieve the same variance as compared to conventional stochastic sampling. Our results show that uncertainty in volcanic ash dispersion forecasts can be straightforwardly calculated and communicated, and highlight the need for the volcanic ash forecasting community and operational end-users to jointly choose acceptable levels of variance for ash forecasts in the future.

Keywords Volcanic ash · Probabilistic forecasts · Confidence intervals · Weather patterns

Introduction

A key consideration in forecasting the spatial extent and intensity of future natural hazard events is accounting for the variability in environmental conditions that influence them. This is commonly investigated using a probabilistic approach, which involves simulating sufficiently many realistic event scenarios, where environmental conditions are sampled from a distribution of sufficient duration, to incorporate much of the likely variability (Bonadonna, 2006; Jenkins et al., 2012). The simulation results are then aggregated to estimate the probability of exceeding a given threshold, or the threshold exceeded at a given probability (Rougier, 2013). In the volcanological literature, there is no commonly applied principle that informs the number of simulations that is 'sufficient' to characterise the variability of environmental conditions, so choices are often made on a pragmatic basis (Bonadonna et al., 2005; Jenkins et al., 2012, 2015; Harvey et al., 2020; Zidikheri and Lucas, 2021). Furthermore, these predictions are often stated without a confidence bound on the prediction, which is critical given that the use of simulations to determine hazard probabilities is an exercise in statistical estimation (Rougier, 2013). In this paper, we explore the relationship between sample size, uncertainty, and sampling strategies in probabilistic assessments of volcanic ash dispersion.

Atmospheric dispersion of volcanic ash is driven by wind fields and atmospheric turbulence; dispersion models are forced using measured or interpolated wind fields (typically reported as means at regular intervals), and turbulence is parameterised within the models. Given a historical record of wind fields, a convenient approach is to sample the wind

Editorial responsibility: S. Barsotti

Jeremy Phillips and Shannon Williams contributed equally to this work.

✉ Jeremy Phillips
J.C.Phillips@bristol.ac.uk

Extended author information available on the last page of the article

fields *stochastically* by randomly selecting the start date in this record (Bonadonna et al., 2005; Jenkins et al., 2012). If the eruption source parameters are held constant to solely investigate the influence of the environmental forcing, then this is termed the 'One Eruption Scenario' (Bonadonna, 2006). The stochastic approach assumes that, with enough randomly selected samples, the variability in the natural occurrence of different wind directions and velocities will be reproduced in the wind field data used and that there are no underlying trends in the wind behaviour which evolve over the long term (i.e. the dataset is stationary). Stochastic sampling is widely used in probabilistic assessments of volcanic ash dispersion (e.g. Connor et al. 2001; Bonadonna et al. 2005; Bonadonna 2006; Macedonio et al. 2008; Jenkins et al. 2012, 2015; Biass et al. 2016). These studies vary widely in the number of individual simulations that are used to construct the ensemble, for example from 73 (Macedonio et al., 2008) to 19,200 (Jenkins et al., 2015), and primarily limitations on computational resources or time are the justifications given for the choice of ensemble size. Quantifying the uncertainty associated with the choice of sample size is crucial to appropriately balance the trade-off between accuracy and computation.

In this paper, we introduce the use of standard statistical measures to compare results from ensembles of different sizes. The use of stochastic simulations to create probabilistic ash dispersion forecasts is guided by the idea that a larger number of simulations will better reflect the range of environmental conditions that control dispersion. To be concrete, we consider estimation of the *expectation* of some output of the simulator, i.e. the average of that output over all possible inputs. A natural estimator for this quantity is the *sample mean*, i.e. the average of that output over a finite number of randomly sampled inputs, requiring that number of simulator runs. The sample mean of only a few outputs is typically quite variable, in the sense that it may be quite different to the true mean, and one way to quantify this variability is through the *variance* of the sample mean: the expected squared deviation of the sample mean from the true mean. In fact, the variance of the sample mean is inversely proportional to the sample size itself, so running more simulations will decrease the variance of the estimator.

A typical aim of stochastic simulation is to make a prediction together with some quantification of the associated uncertainty, and in some cases, it is necessary for the uncertainty to be below some prescribed level (Rougier, 2013). A standard way to accomplish the former is to construct *confidence intervals*, which depend explicitly on the variance of the estimator, and the former can be addressed by drawing sufficiently many samples such that the variance is below the prescribed level. Presenting the results of a stochastic method without an assessment of the uncertainty in them provides incomplete information to end-users, as previously emphasised for the estimation of exceedance probabilities for natural hazards generally (Rougier, 2013) and for ash dispersion ensembles in particular (Marzocchi et al., 2015). Looking forward, guidance for future operational probabilistic forecasts of airborne volcanic ash concentration in the atmosphere requires their results to be stated with associated confidence intervals or variance estimates (ICAO, 2017; WMO-IUGG, 2019).

The use of stochastic sampling of wind fields for volcanic ash dispersion assumes that the wind fields are stationary in time, and so the historical record for wind fields provides an appropriate distribution for wind fields at an unspecified future date. There is, however, an underlying structure to the wind fields over shorter-term intervals because they occur as a result of large-scale weather systems. This structure allows daily conditions to be identified as belonging to one of a set of weather patterns; for example, the weather systems over northern Europe have been classified into a set of 29 synoptic (1000 km scale) weather regimes, named *Grosswetterlagen* (James, 2007). Grouping wind fields into regimes allows an alternative *stratified sampling* approach to be taken (Cochran, 1977): sample forcing data from within individual weather regimes, then weight each resulting set of simulations by the frequency of occurrence of that pattern. The potential benefits of this approach are to reduce the number of simulations needed to reproduce the variability of the natural wind fields, because the number of patterns is smaller than the number of individual different states of the wind field. Stratified sampling approaches are widely used to achieve reductions in variance in other applications (e.g. Hens and Tiwari 2012; D'Amato et al. 2012; Da Silva Fonseca Junior et al. 2015).

In this paper, we will demonstrate the use of variance computations to compare ensemble predictions of volcanic ash dispersion and explore the potential of stratified sampling using weather regimes to reduce the variance in those predictions. We use a dataset of simulations made using the dynamic ash dispersion model `FALL3D` (Folch et al., 2009), to assess potential ash impacts to northwestern Europe from a representative eruption of an Icelandic volcano. The paper is set out as follows: we first introduce the statistical background that quantifies the relationship between the number of samples, the estimates of exceedance probabilities, and the variance of these estimates. We then present the data and methods used, and in the "Results" section, we present results comparing the variances of ensembles of different sizes and demonstrate that stratified sampling of weather regimes reduces the number of samples needed to achieve a desired variance. In the "Discussion" section, we discuss the benefits of quantifying and presenting measures of uncertainty, and of employing a stratified sampling approach, in probabilistic assessment of volcanic ash dispersion.

Statistical background

This section presents the background, principles, and definitions of stochastic sampling in the context of volcanic ash hazard assessment. We introduce the widely used method for estimating the exceedance probability of an ash concentration threshold and demonstrate the calculation of its variance for use in the construction of confidence intervals before proceeding to illustrate the relationship between such probabilities and their variances. We show that the number of samples required for exceedance probability estimates is intrinsically linked to the magnitude of the probability itself and, hence, the threshold of interest. Furthermore, the desire to attain some level of confidence in the estimate suggests the need for some minimum number of simulations, indicating the existence of a relationship between the threshold of interest, the variance of its exceedance probability estimate, and the computational resources required for accurate estimation.

An ensemble is constructed by carrying out a number of simulations whose inputs are sampled stochastically from some distribution which is assumed to be representative of the real world. The outputs returned by the simulator are aggregated to construct the resulting probabilistic assessment of the eruption scenario. Deterministic simulators such as `FALL3D` (Folch et al., 2009) simulate the dispersion and transportation of ash over time given wind field data and eruption source conditions. Since no additional stochasticity is introduced by such a simulator, by keeping the volcanological inputs constant across simulations and drawing wind fields from historical data, we arrive at an ensemble of simulations whose variability is dependent only upon the wind field data. Provided that the historical record is representative of the long-term variability in wind fields, the ensemble should encapsulate the variability of likely real-world outcomes of the eruption scenario.

Drawing such wind field data from a historical dataset consists of choosing a date from the historical record and providing the corresponding wind field data to the simulator. We can therefore view the simulator output as the application of a function to some randomly chosen start date.

Stochastic sampling for exceedance probability estimation

Denoting the set of possible start dates by $\mathcal{Z}$, we view the start date provided to the simulator as a random variable Z which is drawn uniformly at random (i.e. with equal probability) from $\mathcal{Z}$. In the One Eruption Scenario, since the volcanological inputs remain fixed, we then view the output of the (deterministic) simulator as the result of applying some function ϕ to the realised value z of Z, $\phi(z)$. The exceedance probability of an ash concentration threshold c μg m^{-3}, is defined as the probability that, at the location of interest, the maximum ash concentration that persists for more than some specified period of time (e.g. 24 h) is greater than c μg m^{-3}. Mathematically, we can view this maximum concentration as the result of applying a second function ψ to the output of the simulator, which we represent by $\psi \circ \phi(z)$, where $\circ$ denotes the composition of functions.

The random variable $X = \mathbb{1}\{\psi \circ \phi(Z) \geq c\}$, where $\mathbb{1}$ denotes the indicator function, then represents the event of exceedance. X has a Bernoulli distribution with success parameter p, which depends implicitly on ψ, ϕ, and c: X takes value 1 with probability p and value 0 with probability $1 - p$. The expectation of X is $\mathbb{E}[X] = p$, and its variance (its expected squared deviation from its mean) is $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = p(1-p)$. The exceedance probability is precisely p, and we aim to estimate this value with a high degree of confidence.

Typically, a probabilistic ash hazard assessment will estimate the exceedance probability of an impact threshold via a simple Monte Carlo approach, which we hereby refer to as the stochastic sampling approach. Given some specified number of samples n, we sample $Z_1, \ldots, Z_n$ independently and uniformly from the set $\mathcal{Z}$ and transform these into independent Bernoulli samples $X_1, \ldots, X_n$, where $X_i := \mathbb{1}\{\psi \circ \phi(Z_i) \geq c\}$ for $i \in \{1, \ldots, n\}$. We estimate p intuitively by the sample mean of $X_1, \ldots, X_n$,

$$p^n_{\text{simple}} := \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1}$$

where the superscript n indicates the estimate is computed from n samples and the subscript that a simple Monte Carlo approach is used. This provides an unbiased estimate of p, meaning that the expected value of p^n_{simple} is p. Its variance is given by

$$\mathrm{Var}\left(p^n_{\text{simple}}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}p(1-p), \tag{2}$$

which we can estimate by replacing p with its estimate p^n_{simple}; the estimate will tend towards the true value of the variance as n increases. We proceed to describe how to use this variance estimate to obtain an approximate confidence interval for p.

Confidence intervals

Given data $\mathbf{X} := (X_1, \ldots, X_n)$, a confidence interval (CI) for a parameter is a random interval $[L(\mathbf{X}), U(\mathbf{X})]$, where the endpoints $L(\mathbf{X})$ and $U(\mathbf{X})$ of the interval are themselves random variables. A CI should be constructed to have a high probability of containing the true value of that parameter.

The size $U(\mathbf{X}) - L(\mathbf{X})$ of the interval therefore indicates how much uncertainty there is in the value of the parameter. More specifically, for a general parameter θ associated with the distribution of X, a $100(1-\alpha)\%$ confidence interval for θ is an interval $[L(\mathbf{X}), U(\mathbf{X})]$ such that the true value of θ lies within the interval with probability at least $1-\alpha$ (DeGroot and Schervish, 2012):

$$\mathbb{P}(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) \geq 1-\alpha, \tag{3}$$

where $\alpha \in (0, 1)$ is typically small. We call the probability that θ lies between $L(\mathbf{X})$ and $U(\mathbf{X})$ the *coverage* of the CI.

Given data $\mathbf{x}$, we can compute the endpoints $L(\mathbf{x})$ and $U(\mathbf{x})$ of a real, *observed* CI. We construct an asymptotically exact CI for a parameter through the notion of asymptotic normality, which we define formally in Appendix B. The important result to note is that we can construct approximate CIs based on the standard normal distribution given an asymptotically normal estimator. In particular, the sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is an asymptotically normal estimator for the expected value of X, $\mu = \mathbb{E}[X]$, and an accompanying estimator for its variance $\sigma^2 = \mathrm{Var}(X)$ is $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \hat{\mu}_n\right)^2$, so an approximate $100(1-\alpha)\%$ CI for μ is

$$\hat{\mu}_n \pm z_{\alpha/2}\sqrt{\frac{\sum_{i=1}^n \left(X_i - \hat{\mu}_n\right)^2}{n}}. \tag{4}$$

where $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ and Φ^{-1} is the inverse cumulative distribution function of a standard normal random variable. A traditional choice of confidence level is $\alpha = 0.05$ so that $z_{\alpha/2} = 1.96$, giving a 95% CI. Furthermore, since Eq. 1 is simply the sample mean of a Bernoulli(p) random variable, an approximate $100(1-\alpha)\%$ CI for the exceedance probability p is given by

$$p^n_{\text{simple}} \pm z_{\alpha/2}\sqrt{\frac{p^n_{\text{simple}}(1-p^n_{\text{simple}})}{n}}. \tag{5}$$

When probabilities are small and we want to capture and convey the risk of a set of hazardous events, it is often beneficial to view these probabilities on a logarithmic scale. In particular, we want to visualise estimates of probabilities of differing magnitudes, and our confidence in these estimates, on the same scale. In Appendix B, we show that an approximate $100(1-\alpha)\%$ CI for $\log p$ which is centred about $\log\left(p^n_{\text{simple}}\right)$ is given by

$$\log\left(p^n_{\text{simple}}\right) \pm z_{\alpha/2}\sqrt{\frac{1-p^n_{\text{simple}}}{np^n_{\text{simple}}}}. \tag{6}$$

Relationship between threshold and variance

When carrying out a probabilistic hazard assessment, we are usually interested in a set of several impact thresholds rather than a single value. It is clear that the exceedance probability associated with a high ash concentration threshold will be lower than that of a smaller threshold. We would therefore expect more simulations to be required in order to observe at least one exceedance event in our ensemble as the threshold increases. In this section, we demonstrate how to choose our ensemble size such that the lowest exceedance probability of interest can be estimated with some desired level of confidence.

For estimating the value of some p, we must consider a minimal sample size n for which a sensible CI is achievable. Suppose we have an ensemble of size 1000 and the true value of the exceedance probability p for some threshold of interest is 10^{-4}. Then, the closest possible estimates we might obtain for p using Eq. 1 are either 0 or 10^{-3}. The former arises by observing no exceedances in the ensemble and the latter by observing a single exceedance in 1000 samples, providing us with an estimate which is 10 times greater than the truth. Furthermore, approximate 95% CI for p would then be either $(0, 0)$ or $(-9.6 \times 10^{-4},\ 2.96 \times 10^{-3})$, respectively.

A natural way to examine the relationship between the ensemble size n and exceedance probability p is to consider the expectation of the number of exceedances $Y_n := \sum_{i=1}^n X_i$:

$$\mathbb{E}[Y_n] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = np. \tag{7}$$

In order to have $\mathbb{E}[Y_n] \geq 1$, we must set $n \geq 1/p$. As such, if we have some prior belief on p for the highest threshold of interest, we can use this relationship to set the number of samples. If $p = 10^{-4}$, then at least 10,000 samples are needed in order to expect to observe at least one exceedance.

A single exceedance event does not provide us with much confidence in our estimate, however. Having a higher degree of confidence in our estimate corresponds to reducing our variance such that our CIs centred about this estimate are smaller. Guidance exists for choosing n such that the CI has a desired width, such as Liu and Bailey (2002). We note that for estimates of small Bernoulli probabilities p, the associated variances $p(1-p)/n$ are also small, but for small n, these variances are large when compared with the value of p itself. It should therefore be beneficial to consider a measure of the variance in relation to the probability itself rather than the variance in isolation. We introduce the *relative variance* of the estimate as the ratio of the variance of the estimate and its squared expectation. For Bernoulli probabilities, this provides a natural measure of the variability in relation to the

scale of the probability itself:

$$\mathrm{Var}\left(\frac{p^n_{\mathrm{simple}}}{p}\right) = \frac{\mathrm{Var}\left(p^n_{\mathrm{simple}}\right)}{p^2} = \frac{1-p}{np}. \quad (8)$$

We can choose n such that the relative variance is less than some $\varepsilon \in (0, 1)$, giving $n > (1-p)/\varepsilon p$. Given some domain knowledge of the exceedance probability for our highest threshold of interest, we can then decide on a suitable number of simulations such that the relative variance of our estimate is constrained. Notice that in Eq. 6, the term within the square root is an estimate of Eq. 8. Thus, choosing n such that the relative variance is restricted directly reduces the width of the CI for $\log p$.

We proceed to describe an approach for further reducing the variance of our exceedance probability estimates through a straightforward adaptation to stochastic sampling referred to as stratification.

Data and methods

We compare and contrast probabilistic tephra hazard across Europe, and at a number of key locations (Fig. 1 and Table 1), to investigate how, where, and why hazard differs between approaches using stochastic and targeted sampling of meteorological conditions. Tephra from a hypothetical Volcanic Explosivity Index (VEI; Newhall and Self 1982 4 eruption from Iceland is simulated. This is one of the most prominent sources of tephra for northern Europe because of the frequency of Icelandic eruptions and the predominance of westerly winds in the region (Crosweller et al., 2012). Key locations were chosen for analysis here (Fig. 1 and Table 1) because they have record of Icelandic tephra deposits (Swindles et al., 2011) and to cover a range of bearings and distances.

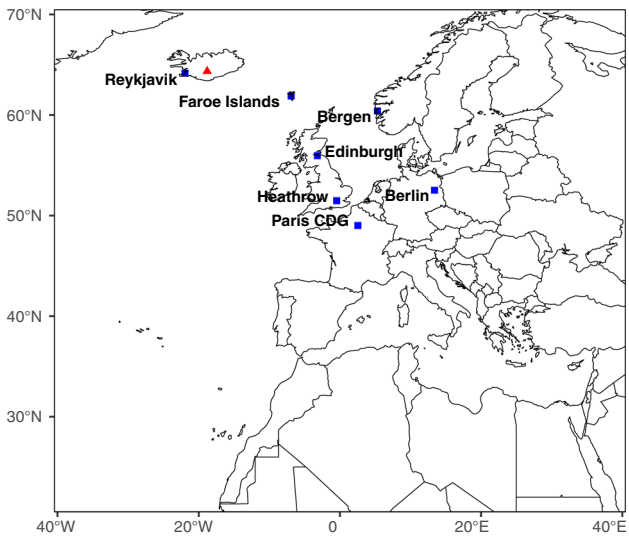

Fig. 1 European sites chosen for testing and the volcanic source (red triangle) in Iceland

Table 1 Distance and bearing (degrees from north) of the European sites in Fig. 1, relative to the volcanic source in Iceland

Location	Distance (km)	Bearing (°)
Reykjavik, Iceland	160	270
Faroe Islands, UK	645	110
Edinburgh, UK	1275	130
Bergen, Norway	1300	100
Heathrow, UK	1800	135
Paris CDG, France	2120	135
Berlin, Germany	2290	110

Volcanic source and ash dispersion modelling

We simulated volcanic ash dispersion from Iceland to northern Europe using the open-source dynamic three-dimensional advection–diffusion `FALL3D` model version 7.0 (Folch et al., 2009). In this application, we provided hourly reanalysis meteorological data as input to calculate ash transport; the model interpolates to hourly outputs of three-dimensional (latitude, longitude, and altitude) ash concentration and ground ash thickness. We simulated 4 days (96 h) of ash transport starting from the onset of an 8 h eruption—this was found to be sufficiently long to capture long-range ash transport and deposition.

We chose volcanic source conditions to be representative of a VEI 4 size eruption as follows. The plume height was chosen based on empirical fits of observed plume heights to erupted volume (Jenkins et al., 2007). Ash within the plume was assumed to follow a Suzuki distribution (Suzuki, 1983), with parameters such that more ash is released beyond the convective thrust region and in the upper portion of the plume (Table 2). Eruptions were considered to last for 8 h to represent a large silicic eruption (Mastin et al., 2009). The erupted total particle sizes follow a Gaussian distribution with ten size classes between 1 and 10 ϕ (0.9 μ and 0.5 mm), with a mean of 2.5 ϕ (0.18 mm) and standard deviation of 4.5 ϕ (0.04 mm), following those derived by Folch et al. (2012) for the 2010 Eyjafjallajökull eruption in Iceland. Particle densities of 1500 kg m^{-3} and 2500 kg m^{-3}, consistent with intermediate magma compositions, are associated with coarse (1 ϕ) and fine (6 ϕ) particles, respectively. The commonly used Ganser model (Ganser, 1993) is used within the model to calculate the terminal settling velocities of the near spherical particles through the atmosphere. The mass flow rate of the eruption (in kg s^{-1}) is calculated within `FALL3D` using

Table 2 Overview of FALL3D model inputs for a VEI 4 eruption

Variable	Model input
Plume height	19 km
Eruption duration	8 h
Erupted volume	0.32 kg m^{-3}
Total mass	28.5×10^6 kg
Plume shape	Suzuki source with parameters $A = 4$ and $\lambda = 1$
Particle size distribution	Gaussian distribution between 1 and 10 ϕ with $\mu = 2.5\,\phi$ and $\sigma = 4.5\,\phi$
Particle density	1500 kg m^{-3} (coarse particles) 2500 kg m^{-3} (fine particles)
Particle sphericity	0.9
Diffusion coefficients	5000 m s^{-2} (horizontal) 500 m s^{-2} (vertical)
Meteorological data	96 h of reanalysis data from ECMWF ERA-5

an empirical fit between mass flow rate and plume height due to Mastin et al. (2009), assuming a magma density of 2500 kg m^{-3}. Parameters used in the dispersion modelling are summarised in Table 2.

For this study, we used a 'One Eruption Scenario' approach (Bonadonna, 2006) and compiled a dataset of 6000 simulation results using these initial conditions with meteorological forcing data sampled from 6-hourly mean data from the ECMWF ERA-5 dataset, where start dates were randomly selected between January 1997 and December 2005.

Meteorological data

The vast majority of large-scale weather variability can be classified into a manageable number of synoptic patterns or regimes. A weather regime is any configuration of the weather that tends to remain relatively constant for a period of a few days to weeks (James, 2006); they may be classified manually or using an objective classification system. Weather regimes have been classified for the UK (Jones et al. 1993, $n = 8$), Europe and the north-east Atlantic (James 2006, $n = 29$), the Western US (Robertson and Ghil 1999, $n = 6$), the Northern Hemisphere (Barnston and Livezey 1987, $n = 13$), South America Solman and Menendez 2003, $n = 5$), and New Zealand (Kidson 2000, $n = 2$, across four broad types), among others. The scale over which regimes are categorised therefore varies significantly from the relatively small island of the UK to a whole hemisphere. For an Icelandic eruption, we have used the Grosswetterlagen regimes described in the following section.

Grosswetterlagen system

The widely used Grosswetterlagen (GWL) series of synoptic weather regimes is the only classification system currently in use that can capture both large-scale and local weather regime characteristics (James, 2006). The GWL system was developed by Baur et al. (1944) and is now maintained by the German weather service (DWD, 2015). More recently, James (2007) produced an objective method for classification that can be used with either the ECMWF ERA-5 or NCEP/NCAR reanalysis data, providing GWL classifications from 1948 to present. We obtained a dataset of GWL regimes assigned for every daily ECMWF ERA-5 reanalysis record between January 1997 and December 2005 (Parker, Pers. Comm., 2014). The ECMWF ERA-5 reanalysis catalogue contains global meteorological records at a spatial resolution of 0.25°, approximately 28 km at the equator, for 137 hybrid sigma-pressure levels, related to altitudes up to more than 80 km above sea level (dataset available here).

GWL regimes typically last 4 days in duration, although some regimes (WZ, SWZ, NEA, SEZ) are as short as 3 days and some (WZ, WA, BM, NWZ, SWA, HNA, HNFZ) as long as a week or more (Fig. 2). Descriptions for each GWL regime are provided in Appendix A. Westerly regimes are the most prevalent, although a zonal ridge of high pressure across Central Europe (BM), similar to the weather patterns that dispersed ash towards mainland Europe during the Eyjafjallajökull crisis in 2010, has a relatively high probability of occurrence.

In our analysis, meteorological inputs are chosen randomly from the ECMWF catalogue between 1997 and 2005 and provided to the simulator as 96 h of reanalysis data (Table 2). As GWL patterns last 4 days on average, it is likely that a 4-day period chosen at random from the catalogue will contain a transition between GWL regimes. In order to simplify the process of stratified sampling (which we introduce in the following section) via GWL regimes and to ensure that the resultant number of strata (groupings of the data) is not excessively large, we classify each 4-day period as belonging to the GWL classification of the first day, regardless of any transition between regimes occurring during that period.

Stratification for variance reduction

GWL regimes represent a partitioning of the whole set of daily wind fields into a smaller number (29) of groups with distinctive properties. Given that this grouping is possible, we make use of a statistical sampling technique called *stratification* to reduce the variance of an estimate of an expectation (Cochran, 1977). In our case, each GWL regime constitutes a separate stratum, and samples can be drawn independently from each stratum to form estimates which are then combined to obtain an unbiased estimate of the expectation whose

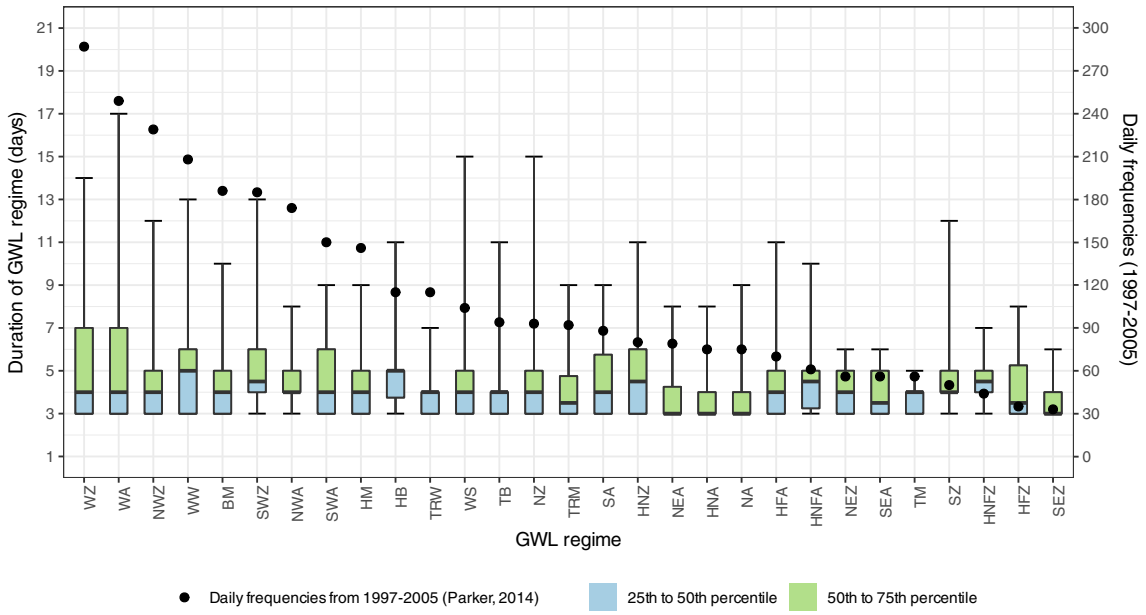

Fig. 2 Duration distributions for each of the 29 GWL regimes. Boxplots show the minimum, 25th percentile, median, 75th percentile, and maximum durations recognised for each regime across the 1997 to 2005 ECMWF ERA-40 reanalysis dataset. Regimes are ordered from most frequently observed (WZ) to least (SEZ), and the daily frequencies between 1997 and 2005 are shown by the black dots

variance is lower than that of the comparable stochastic sampling estimate.

Given that we can assign an individual GWL regime to each start date in the ECMWF ERA-5 reanalysis record, we can determine the probability of occurrence of an individual regime or its *weighting* within the distribution. More formally, for the 29 GWL regimes, we divide the set of start dates $\mathcal{Z}$ into $J = 29$ disjoint regions $\mathcal{Z}_1, \ldots, \mathcal{Z}_J$, according to the GWL classification for that date. Then, the following 96 h of meteorological data provided to the simulator is regarded as belonging to that GWL classification. We then refer to $\mathcal{Z}_j$ as the jth *stratum*. If the random variable Z is drawn uniformly at random from $\mathcal{Z}$, the probability that Z was drawn from stratum $j \in \{1, \ldots, J\}$ is the *weight* of stratum j, denoted by

$$w_j := \mathbb{P}\big(Z \in \mathcal{Z}_j\big), \tag{9}$$

where the sum of these probabilities is one: $\sum_{j=1}^{J} w_j = 1$.

The key to stratification is the fact that the population mean is the weighted sum of stratum means (Cochran, 1977),

$$p = \sum_{j=1}^{J} \mathbb{P}\left(Z \in \mathcal{Z}_j\right) \mathbb{P}\left(X = 1 \,\middle|\, Z \in \mathcal{Z}_j\right) \tag{10}$$

$$= \sum_{j=1}^{J} w_j p_j, \tag{11}$$

where, for each $j \in \{1, \ldots, J\}$, p_j is the exceedance probability associated with stratum j:

$$p_j := \mathbb{P}\left(X = 1 \,\middle|\, Z \in \mathcal{Z}_j\right). \tag{12}$$

Stratified sampling estimates

To carry out a stratified sampling procedure, we must know the weights $w_1, \ldots, w_J$ and be able to sample directly from each stratum. In our case, we have access to the ECMWF ERA-5 catalogue from 1997 to 2005 alongside the GWL classification for each date in the catalogue. We can therefore calculate the weights using the relative frequencies of occurrence of each GWL regime and sample uniformly from each stratum by choosing a start date for the simulation from the set of dates with the corresponding weather classification. For each stratum $j \in \{1, \ldots, J\}$, we sample n_j start dates $Z_{j,1}, \ldots, Z_{j,n_j}$ independently and uniformly at random from stratum j and consequently obtain n_j independent Bernoulli samples $X_{j,1}, \ldots, X_{j,n_j}$, where $X_{j,i} := \mathbb{1}\left\{\psi \circ \phi\left(Z_{j,i}\right) \geq c\right\}$ for $i \in \{1, \ldots, n_j\}$. We estimate p_j by the jth stratum sample mean:

$$p_j^n := \frac{1}{n_j} \sum_{i=1}^{n_j} X_{j,i}, \tag{13}$$

which is an unbiased estimator of p_j with variance $p_j(1 - p_j)/n_j$. The *stratified sampling estimate* of p is the

weighted sum of the estimates of p_j:

$$p^n_{\text{strat}} := \sum_{j=1}^{J} w_j p^n_j. \tag{14}$$

This is unbiased with variance given by

$$\text{Var}\left(p^n_{\text{strat}}\right) = \sum_{j=1}^{J} \frac{w_j^2}{n_j} p_j(1-p_j), \tag{15}$$

which we can estimate by replacing p_j with p^n_j.

Proportional stratum allocation

Proportional allocation (Cochran, 1977) is a straightforward way to assign the values $n_1, \ldots, n_J$. For some positive m, let $n_j = \lceil m w_j \rceil$, i.e. the smallest integer greater than or equal to $m w_j$. The total number of samples is $n = \sum_{j=1}^{J} n_j \leq m + J$.

The estimator is very similar to the standard Monte Carlo estimator Eq. 1, except that the number of samples in each region $\mathcal{Z}_j$ is deterministic rather than random. We obtain the variance bound

$$\text{Var}\left(p^n_{\text{strat}}\right) \leq \frac{1}{m} \sum_{j=1}^{J} w_j p_j(1-p_j), \tag{16}$$

with equality if $n = m$. To compare stratified sampling with regular stochastic sampling, it is helpful to decompose the variance of $X \sim \text{Bernoulli}(p)$ into within- and between-stratum variances as follows:

$$p(1-p) = \sum_{j=1}^{J} w_j p_j(1-p_j) + \sum_{j=1}^{J} w_j (p_j - p)^2. \tag{17}$$

By comparing Eq. 16 with Eq. 2, stratification is likely to be beneficial if the within-stratum variances are dominated by the between-stratum variance, since the between-stratum variance is not present in Eq. 2. That is, if samples drawn from the same stratum are typically similar to each other, and samples from different strata are typically distinct. An intuitive explanation for this phenomenon is that by using the precise stratum weights in Eq. 14, one source of uncertainty about p is removed and variability arises only from the variability of estimates of the stratum probabilities p_j.

In our results, we illustrate that proportional allocation allows us to obtain low-variance estimates of exceedance probabilities for a range of thresholds. The results indicate that stratified sampling allows us to achieve results comparable to stochastic sampling from fewer samples, thus reducing the amount of computational resources required to obtain high-precision estimates of exceedance probabilities.

In Appendix C, we describe an optimum stratum allocation which will always reduce the variance of the estimate, where the stratum sample sizes are allocated proportionally to the weights and stratum variances if these values are known. We describe also a variant of stratified sampling referred to as post-stratification (Jagers et al., 1985), whereby the variance of an estimate can be reduced post-hoc if samples have already been drawn according to random sampling, and illustrate an example application of this method.

Results

In this section, we present the results of our analysis. We first present results related to simple Monte Carlo sampling of ensemble members, for common ash dispersion metrics as well as exceedance probabilities, and proceed to compare the results for exceedance probabilities with those obtained via stratified sampling of GWL regimes.

Relationship between sample size and confidence for ash dispersion metrics

Volcanic ash dispersion simulations are typically used to predict ash thickness deposits, and more recently estimates of airborne or ground-level ash concentration have become important for impacts to aviation and infrastructure (Biass et al., 2014; Capponi et al., 2022; Harvey et al., 2020). In this context, probabilistic simulation will allow us to estimate the expected value of these quantities via the sample mean and compute an approximate CI using Eq. 4.

Figure 3 maps the estimated mean ash thickness from a VEI 4 eruption from Iceland, together with 95% CIs, for three ensembles of increasing sample size. Each `FALL3D` simulation took 3 h to complete on average, run in parallel on a high-performance computing cluster. The mean thickness is the cumulative value of ash thickness per unit area after the duration of the simulation, divided by the number of hours (96 h here); we estimate thicknesses in excess of 1 cm proximally to less than 0.1 mm distally. The thickness is shown on a logarithmic colour scale to visualise the orders-of-magnitude difference in deposition over very long transport distances; the CI width colour scale is the difference in logarithm thickness between the upper and lower limits of the 95% CI, and we use it here to illustrate relative differences in confidence between different ensemble sizes.

The variation of mean ash thickness decreases with increasing ensemble size—this can be seen from both the CI limit maps and most easily from the CI width maps. For the smallest ensemble (50 samples), there are orders of magnitude difference in mean ash thickness across northern Europe between the upper and lower limits of the CI, but this difference is significantly reduced if ensemble size is increased to

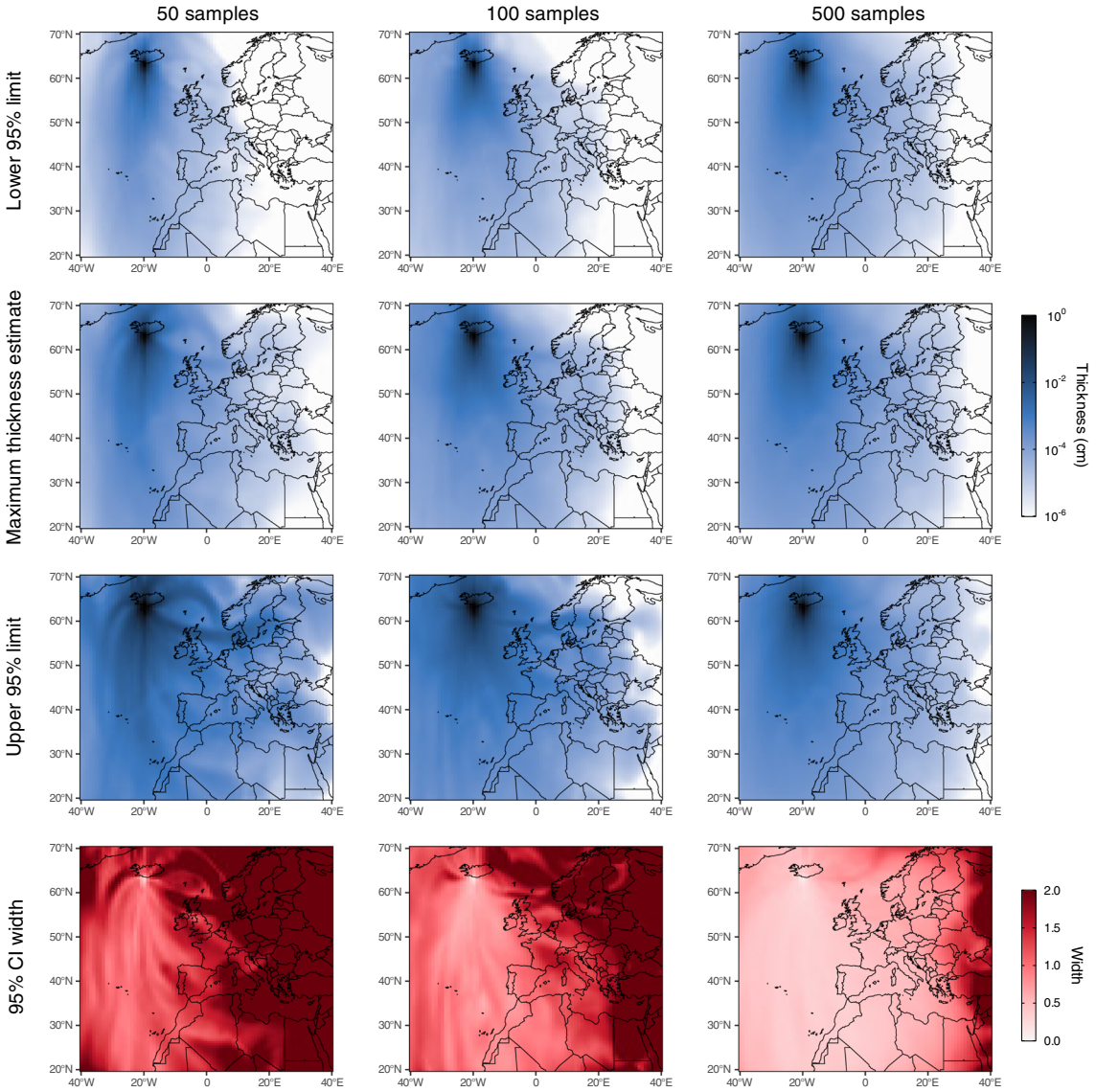

Fig. 3 The variation of mean ash thickness estimates with ensemble size for simulations of a VEI 4 eruption from Iceland. Ensemble size increases from left to right, and each column shows ash thickness maps for the estimated maximum thickness ('maximum thickness estimate') and the lower and upper limits of the 95% CI for the maximum thickness ('lower 95% limit' and 'upper 95% limit', respectively). Note the logarithmic colour scale for thickness in cm. The '95% CI width' maps indicate the size of the CI, which is the difference in logarithm thickness values between the upper and lower limits. Ash thickness is computed from ash deposition using the particle densities given in Table 2

500 samples. For this largest ensemble, there is only a very small difference between the upper and lower ends of the CI, except in the most distal locations (thousands of kilometres from the source). The variation in ash thickness is generally lowest at proximal locations and increases with distance from the source; in proximal locations, ash deposition is dominated by larger particle sizes whose dispersion is less affected by wind field variations because of their higher terminal fall velocities and thus lower residence times in the atmosphere.

In Fig. 4a, we show plots of the variation in estimated ground-level ash concentration at some location as a function of time for ensembles of increasing size. Here, the ash concentration is smoothed over a 24-h window, so that the concentration at x hours from the eruption onset is the average from time x to $x + 24$. This is an important measure for assessing ash impacts to human health, visibility, and infrastructure using filters such as generators and air conditioning systems (Jenkins et al., 2015). On each sub-figure, the mean for the whole set of 6000 samples provides a visual guide as

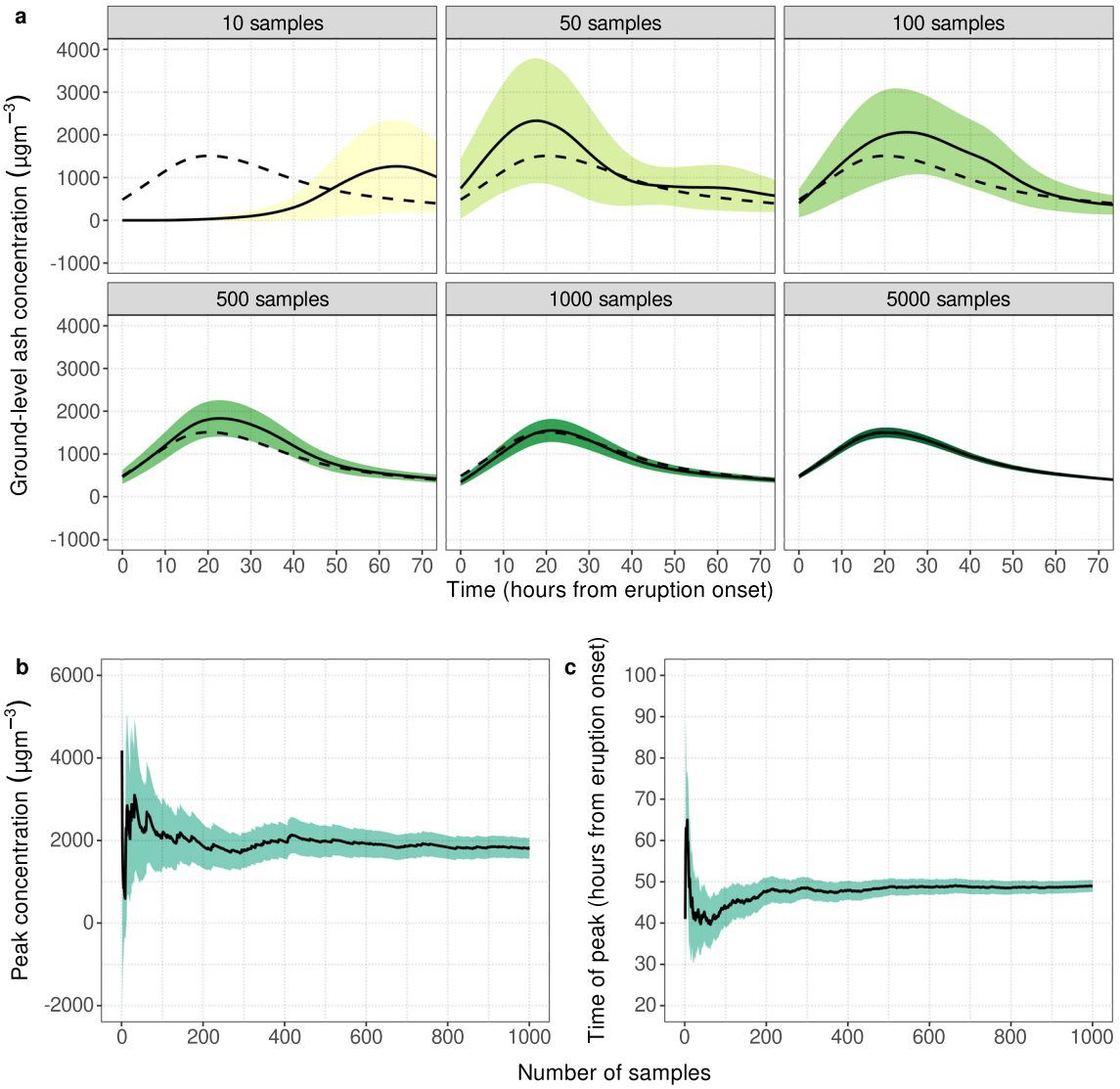

Fig. 4 **a** Ground-level ash concentration estimates over time at Edinburgh, smoothed over a 24-h window, shown for different ensemble sizes. Full lines and shaded regions represent the sample means and 95% CIs, respectively, obtained using the specified number of samples. Dashed lines show the mean obtained from the whole set of 6000 samples. **b** Cumulative estimate (sample mean) of the peak concentration at Edinburgh, against ensemble size, with 95% CIs. **c** Cumulative estimate of the time from eruption onset at which the peak ground-level ash concentration is attained, against ensemble size, with 95% CIs

to how closely the means for different sample sizes match the mean of the whole set, which is our closest estimate of the 'true' value.

The location of interest in Fig. 4 is Edinburgh, 1270 km from the volcanic source. Results using an ensemble of 10 samples (Fig. 4a) show poor agreement with the results of the whole sample set, with the peak mean concentration occurring at 64 h from the eruption onset, compared to the 'true' peak at 20 h. If only a very small sample size is used, it is unlikely that the ensemble will adequately reproduce the dominant trend of ash transport to the UK by northwesterly winds. This particular ensemble is formed of simulations that have resulted in a longer arrival time compared to the mean of the full set of 6000 samples. If a different set of 10 samples were chosen, it is highly likely that the mean of the concentration values at each time, and hence the arrival times, would be significantly different. With a larger ensemble (50–100 samples), the peak timing is closer to that of the whole sample size, and the confidence intervals include the whole sample mean, but span a large range. Ensembles with greater than 1000 samples show good agreement with the 'true' variation and have progressively narrower confidence intervals. We emphasise that each sub-figure is obtained by averaging a set of concentration values and that a different set of the same

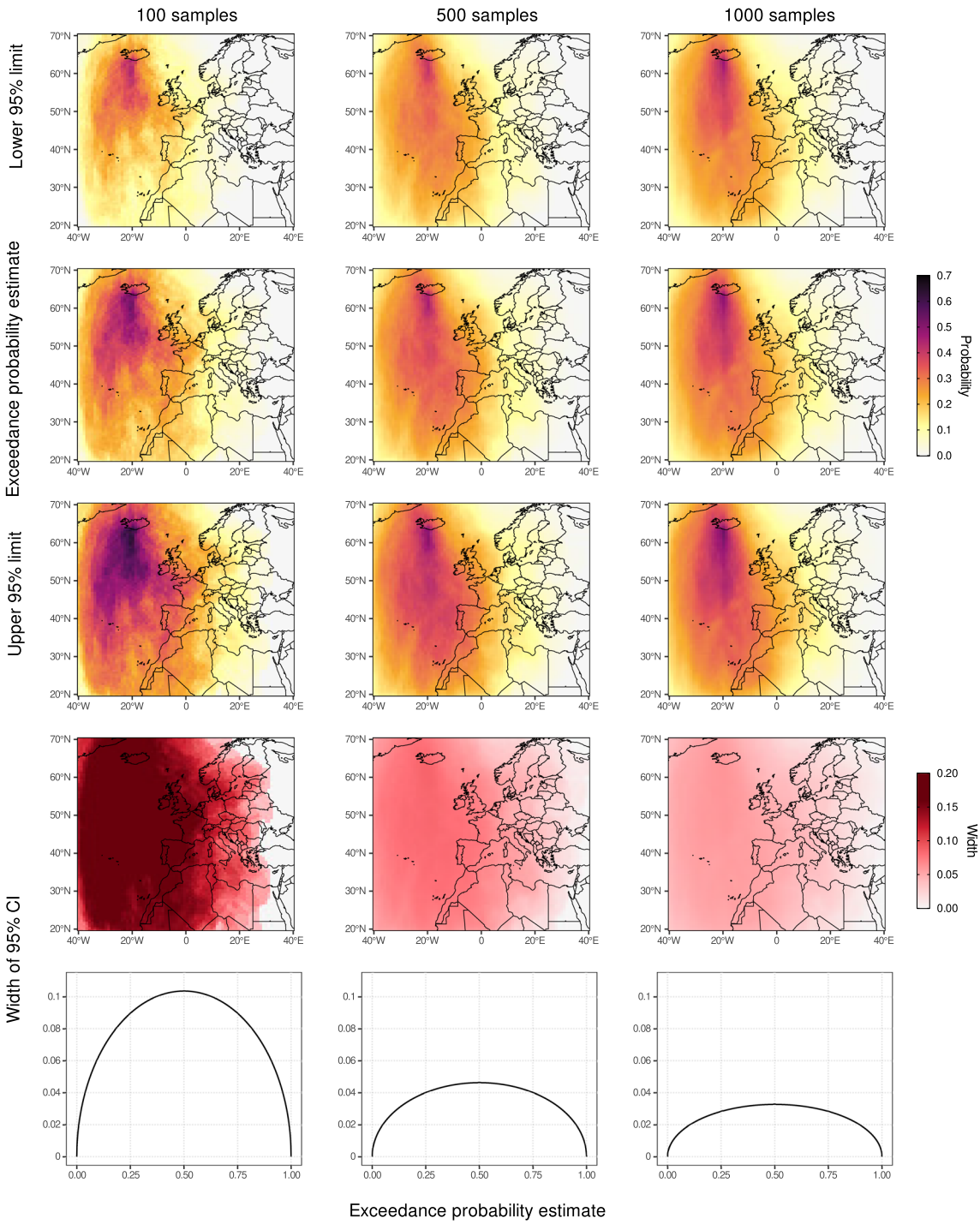

Fig. 5 The variation of exceedance probability estimates with ensemble size for simulations of a VEI 4 eruption from Iceland. The probability of the ground-level ash concentration exceeding 500 μg m^{-3} for 24 h or more is estimated at each point for increasing ensemble size (left to right). Each column shows maps of exceedance probability estimates ('exceedance probability estimate') and the lower and upper limits of the 95% confidence interval for the probability estimate ('lower 95% limit' and 'upper 95% limit', respectively). The 'width of 95% CI' maps illustrate the size of the confidence interval or the difference between the upper and lower limits, and the accompanying curves (bottom row) illustrate these widths as a function of the exceedance probability itself

size would result in a slightly different trajectory, but these differences become smaller as the ensemble size increases.

Figures 4b and c provide cumulative estimates for the peak concentration and the time at which the peak occurs, along with 95% CIs, to illustrate how these CIs shrink with linearly increasing sample size. Note that the mean of the peak concentration is not the same as the peak of the mean concentration (and similarly for peak times), so these estimates do not match the peaks in Fig. 4a.

Relationship between sample size and confidence for exceedance probabilities

For volcanic ash impact studies, the results of stochastic simulations are typically presented using exceedance probabilities for prescribed ash concentration or mass accumulation thresholds (Titos et al., 2022; Jenkins et al., 2022), for which confidence intervals can be computed. Figure 5 maps the probability that the ground-level ash concentration exceeds a value of 500 μg m^{-3} for 24 h or more for different ensemble sizes. As the number of samples increases, the variance is reduced in both proximal and distal locations relative to the source.

Figures 6 and 7 illustrate the relationship between CI widths and threshold for specific locations at different distances from the volcanic source. The exceedance probabilities are calculated using the full ensemble of 6000 simulations, so the lowest probability that can be represented is 1/6000 or about 1.67×10^{-4}. As the threshold increases, the estimated exceedance probability Eq. 1 decreases; then the associated variance Eq. 2 also decreases, and hence the CIs shrink (most easily seen in Fig. 6b). Figure 7 shows how the variance Eq. 2 and relative variance Eq. 8 of the probabilities vary with threshold for the same locations. As the variance of the estimator decreases at a slower rate than the exceedance probability estimator itself, the relative variance increases with threshold and dramatically increases at around 1000 μg m^{-3} (Fig. 7b), where probabilities get very small (Fig. 6b).

Figure 8 displays hazard curves for each location in Table 1, where exceedance probability is shown as a function of threshold on a double-logarithmic scale. These plots clearly show how CI width increases relative to the exceedance probability itself as it decreases, and these CIs become masked when the hazard curve appears to become vertical, as the estimate of the exceedance probability decreases rapidly towards zero due to the finite number of samples.

Given the straightforward calculation of CIs for exceedance probabilities described, we can also calculate how many samples are required in order to restrict the expected size of the CI. This is illustrated in Fig. 9, where we show the minimum number of samples needed to expect the width of the 95% CI to be 50% of the exceedance probability itself (calculated through using the estimates obtained for each threshold in Fig. 6). We note that the number of samples needed increases with threshold (and hence decreasing exceedance probability) and also with distance from the volcanic source. For example, for the width of the 95% CI for the probability of exceeding an ash concentration of 10^4 μg m^{-3} to be 50% of the exceedance probability itself, we require approximately 450 samples for the proximal Reykjavik location and approximately 37,000 samples for the distal Berlin location.

The effect of stratified sampling on estimates

In the context of probabilistic volcanic ash dispersion, stratified sampling aims to reduce the variance in an estimate compared to that made with the same number of stochastic samples or to obtain an estimate with the same variance as an estimate constructed through stochastic sampling but using fewer samples. In this study, we stratify our meteorological input data using GWL regimes. In Fig. 10, we see how exceedance probabilities vary for different GWL regimes with threshold. Notice that there is greater variability in exceedance probability between regimes at lower thresholds because of the different directions of wind fields that make up each regime.

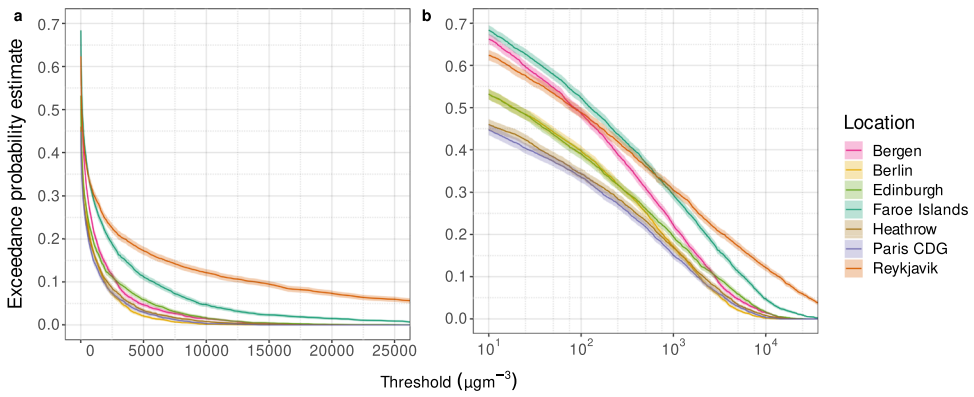

Fig. 6 **a** Exceedance probability estimates against threshold for each location in Table 1, with shaded regions indicating 95% confidence intervals. **b** Exceedance probability estimates against threshold on a logarithmic scale for the same locations, with 95% confidence intervals

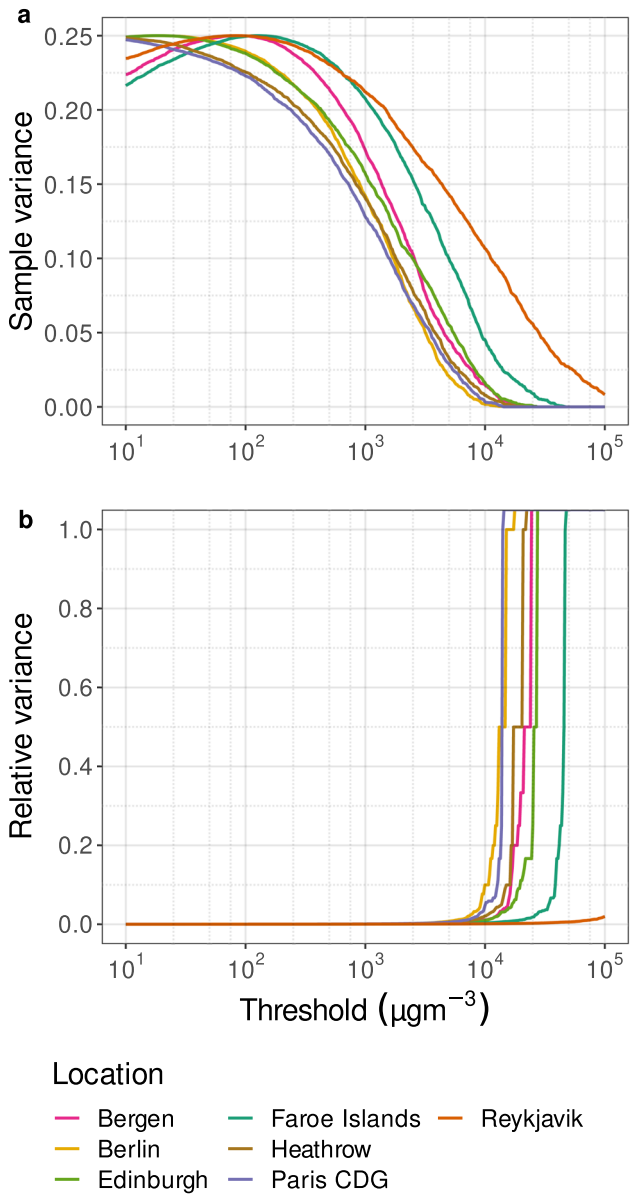

Fig. 7 **a** Sample variance of exceedance against threshold for each location in Table 1. **b** Relative variance against threshold for the same locations

Figure 11 shows, as a percentage, the difference in variance of exceedance probability estimates achieved through the use of proportional stratified sampling using GWL regimes, as compared to stochastic sampling. For all locations and thresholds, proportional sampling from weighted GWL regimes leads to a reduction in the variance of the estimate. The greatest reduction in variance is seen for the lower thresholds, and this benefit decreases with increasing threshold as the exceedance probabilities are lower. We obtain an indication of the percentage reduction in ensemble size that would be required to obtain the same (or lower) variance as the stochastic sampling estimate. For example, for thresholds below 1000 μg m^{-3}, 5 to 17% fewer samples in the ensemble of weighted GWL regimes (i.e. about 5000 to 5700 samples) will give an estimate with the same or lower variance as the set of 6000 stochastic samples. For the FALL3D simulations presented in this study (Table 2), this corresponds to saving up to approximately 1000 h of computational time.

Discussion

In this paper, we have introduced statistical methods for quantifying the uncertainty of probabilistic volcanic ash hazard assessments that are widely used to inform operational decision-makers and risk managers. We have used these approaches to show that using synoptic scale weather regimes, where available, can reduce the number of simulations needed to compute exceedance probabilities to a required degree of confidence, compared to purely stochastic sampling of the wind fields.

When estimating the value of an expectation, a confidence interval can be straightforwardly calculated from the sample mean and variance and the number of samples. We have demonstrated how the use of confidence intervals can enable comparison of the results of ensemble simulations of different sizes. Given that this is not computationally intensive and does not require additional simulations to be performed, we recommend that computing and presenting confidence intervals becomes standard practice when communicating the results of probabilistic volcanic ash hazard assessments. This information provides a measure of uncertainty in the forecast results that can be used by decision-makers, and expressing results with confidence intervals allows the uncertainty in hazard estimation to be propagated through the full risk equation, where typically hazard is considered without any uncertainty. We have shown results for the calculation of confidence intervals for metrics of volcanic ash hazard and for exceedance probabilities and provide an R package for making these calculations whose link is provided in the Supplementary Material section at the end of the paper. These tools can also be straightforwardly implemented within or applied to the results of software packages that make probabilistic volcanic ash calculations, such as TephraProb (Biass et al., 2016). While we have focused on the One Eruption Scenario and explored stochastic sampling of the forcing wind field only, the methods can be applied in exactly the same way to varying eruption source parameters (the Eruption Range Scenario; Bonadonna 2006). When considering the effects of the wind and the source parameters together, we expect the confidence intervals would typically be larger than we have shown here for the One Eruption Scenario (Fig. 6), reflecting the increased variation in the controls on ash dispersion.

Our presentation of the statistical background to confidence interval calculation highlights some important conse-

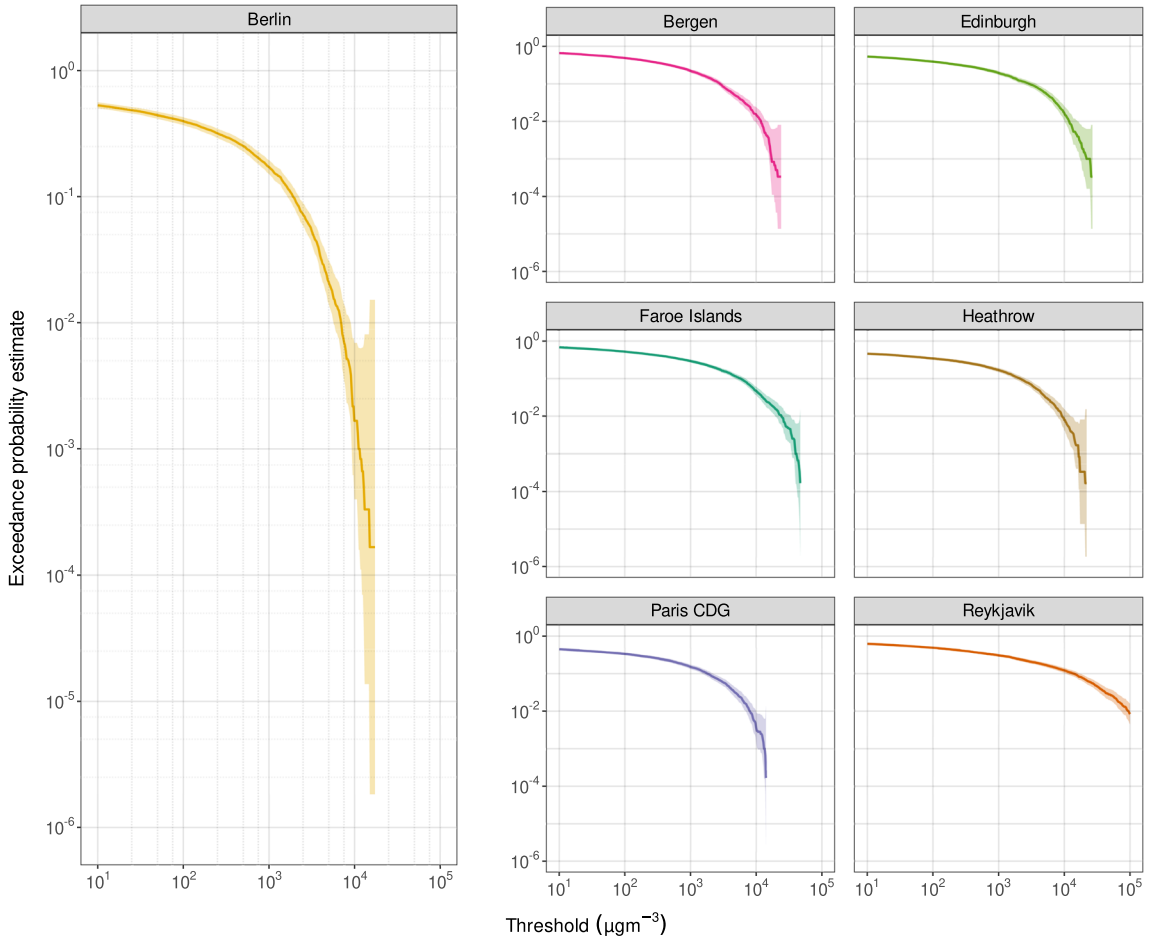

Fig. 8 Hazard curves showing exceedance probability estimates for each location in Table 1 on a double-logarithmic scale, with shaded regions indicating 95% confidence intervals

quences for ensemble design. The exceedance probability estimates that can be computed are directly related to the number of samples in the ensemble. For example, if the ensemble size is 1000, the smallest exceedance probability that can be resolved is 1/1000 or 0.1%. This resolution may be sufficient for some applications, such as assessment of ash impacts to inhabited areas or agricultural land, but may not be for potential impacts to critical infrastructure, which can require exceedance probability thresholds of 10^{-4} (0.0001%) or lower (ONR, 2020). Recognising this relationship between ensemble size and the exceedance probability that can be resolved provides a first step in ensemble design: is the ensemble size sufficient to provide the information needed by decision-makers? Furthermore, the confidence interval for an estimate reduces with increasing sample size (Figs. 4 and 9), so it is also possible to design an ensemble with the number of samples matched to a particular confidence interval range. For example, if repeated laboratory testing identified that an item of equipment failed within a given time period if its exposure to volcanic ash exceeded some concentration threshold, one could specify the number of samples required for an ensemble forecast to calculate the probability of exceeding this threshold for that time period to a desired level of confidence. We provide in Appendix B an accessible and minimal set of statistical results that justify the methods used in this paper. While not touched on in this paper, further improvements to ensemble size selection could be achieved through consideration of stopping rules for stochastic sampling. These require computation of a stopping criteria at each step, or every number of steps (Ata, 2007; Gilman, 1968), and hence facilitate a more complicated sampling process with further computations.

In many cases, it could be helpful for decision-makers to be able to identify where in ensemble simulation results there is greater or lower uncertainty in those results. Confidence interval maps such as presented in Fig. 5 provide an important visual indication of the variability within an ensemble of a given size, as well as quantitative information about the confidence of estimated ash thickness. The confidence interval width plots presented in Fig. 5 provide an immediate indication of regions for which variability of estimates is greatest (further from the volcanic source and in directions

different from the dominant wind direction) and also how the variability changes with ensemble size.

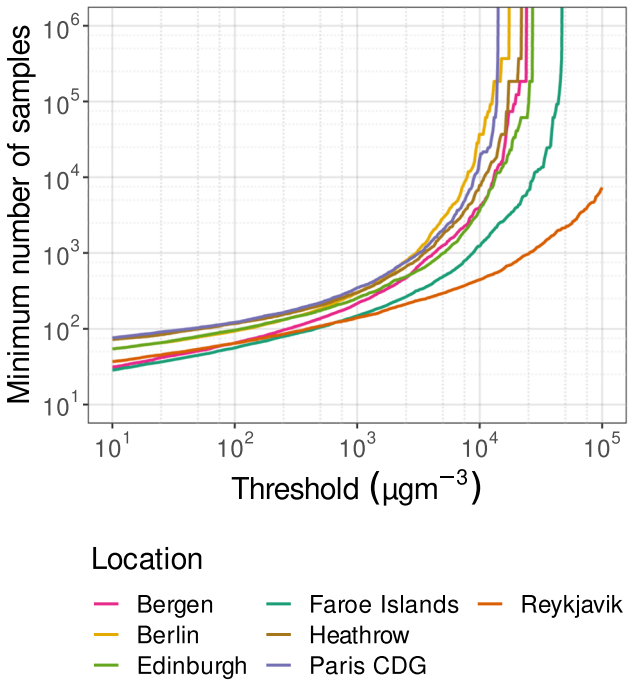

Fig. 9 The minimum number of samples needed for the expected width of the 95% confidence interval of the exceedance probability for a threshold to be 50% of the exceedance probability itself for each location in Table 1

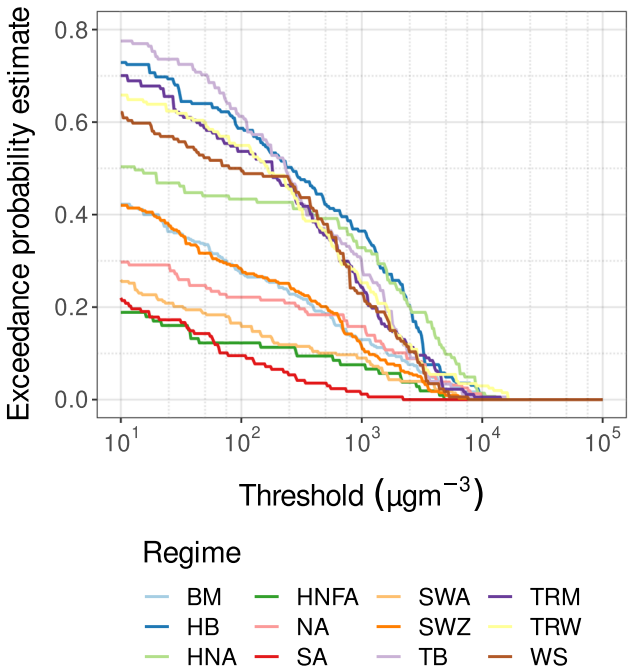

Fig. 10 Exceedance probability estimates against threshold for select (12 out of 29) GWL regimes at Heathrow airport. An ensemble size of 6000 is used

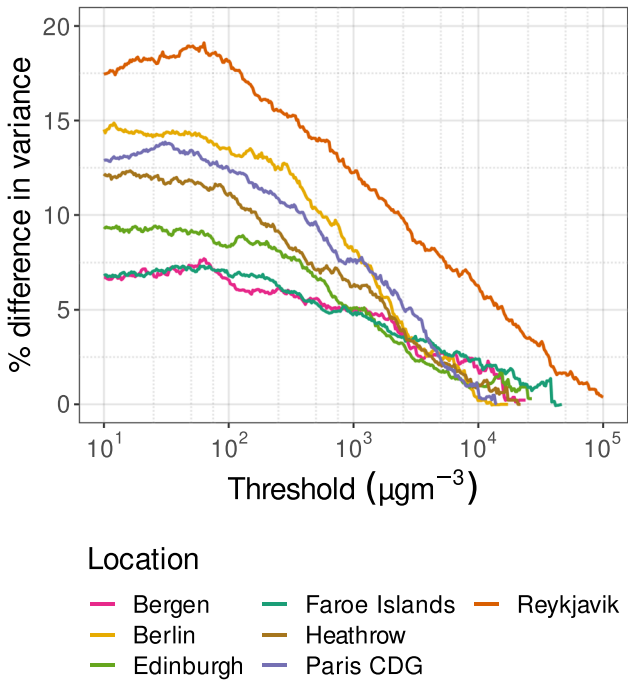

Fig. 11 Difference in variance of exceedance probability estimates from using proportional stratified sampling with GWL regimes compared to stochastic sampling, with the same location, expressed as a percentage. Differences are shown for each location in Table 1, and an ensemble size of 6000 is used

We also use the variance of an estimate to quantitatively compare results from two sets of simulations, in this case between wind fields sampled stochastically and those sampled from wind fields corresponding to particular synoptic-scale weather patterns. The purpose of this example of *stratified sampling* was to investigate whether the use of weather patterns could increase the efficiency of probabilistic ash hazard assessments by reducing the number of samples required in an ensemble to compute an exceedance probability at a given confidence interval. For our simulations over scales of 1000 km, we find that proportional stratified sampling of wind fields corresponding to GWL regimes reduces the variance of estimates, and hence the required number of samples, by 5 to 20% for typical ground-level ash concentration thresholds. Although modest, this reduction could represent a significant saving of computational time when using time-dependent simulators. The ability to use a weather pattern approach relies on allocating each day in the wind field record to a particular regime, which may still need to be done for weather regimes used in other global settings.

The reduction in the number of samples required to compute an exceedance probability with a given degree of confidence using proportional stratified sampling of GWL regimes is achieved because the variance of the stratified sampling estimator is bounded above by the weighted sum of within-stratum variances, which we expect to be dominated by the between-stratum variances when samples from

the same stratum are typically similar to each other. Each set of dispersion simulations for a particular regime could be used to calculate exceedance probability estimates of ground-level ash concentration or ash deposit thickness for a given size eruption occurring into a particular weather pattern. This set of simulations could be used to inform future preparedness and to anticipate the impacts of an event in real-time. In addition, post-stratification can be used to compute approximations with reweighted strata to better reflect environmental conditions for a specific eruption.

We have focused on the estimation of expectations and in their confidence intervals. This is partly to avoid confusion with other sources of uncertainty. In particular, the results presented do not provide information about the variability arising from the capability of the model used. For example, in Fig. 4a, the shaded region indicates uncertainty associated with the sample mean of the ash concentration at each point in time, which is due to the number of samples. With 5000 samples, there is very little uncertainty about the mean, but there is variability in ash concentrations themselves that cannot be quantified by looking at its expected value. Plainly, the quantity being presented is not the 'true' expected value of the quantity given the eruption event in the real world, but that of the quantity output by the simulator given some parameterisations. It is simple to reduce the width of an estimate's CI further by increasing the ensemble size, but this will not reduce the inherent bias which arises from using an imperfect model. In operational environments, the results obtained from using the methods in this paper should always be presented with the caveat that these are based on model assumptions and that 'true' CIs will be larger than the numerical values obtained due to unmodelled variability.

We finally note that, although we used a 95% confidence level in our examples the methods presented in this paper extend to any choice of confidence level. A 95% confidence level is fairly traditional in situations where data is scarce or expensive to collect, as is usually the case in probabilistic volcanic hazard assessment. In practice, choice of confidence level should be context-specific and determined by the application and data. We note also that specialised corrections exist for CIs for Bernoulli probabilities that achieve marginal improvements upon the coverage obtained by the commonly used CI presented in this paper. In operational environments where ensemble size is large enough (typically $n > 40$; Brown et al. (2001)), it can be beneficial to instead use the modified version of Eq. 5 presented in Agresti and Coull (1998). For smaller ensembles ($n < 40$), the Wilson interval may be recommended instead due to its higher coverage (Wilson, 1927; Brown et al., 2001). These minor corrections will result in small improvements in the coverage for an exceedance probability CI.

The use of confidence intervals to compare probabilistic ensembles highlights the need to consider ensemble design across different types of volcanic ash forecasts and identify consistent standards for their presentation to unify probabilistic volcanic ash hazard assessment practice and communicate variability in forecasts. As a starting point, the volcanology community, in collaboration with operational end users, needs to choose the confidence intervals for different forecast applications such as airborne ash concentration or deposited ash thickness.

Appendix A: GWL regime descriptions

Table 3 GWL regime abbreviations and their brief descriptions (James, 2007)

Abbreviation	Description
WA	Anticyclonic westerly
WZ	Cyclonic westerly
WS	South-shifted westerly
WW	Maritime westerly, block Eastern Europe
SWA	Anticyclonic south-westerly
SWZ	Cyclonic south-westerly
NWA	Anticyclonic north-westerly
NWZ	Cyclonic North-Westerly
HM	High over central Europe
BM	Zonal ridge across central Europe
TM	Low (cut-off) over central Europe
NA	Anticyclonic northerly
NZ	Cyclonic northerly
HNA	Icelandic high, ridge over central Europe
HNZ	Icelandic high, trough over central Europe
HB	High over the British Isles
TRM	Trough over central Europe
NEA	Anticyclonic north-easterly
NEZ	Cyclonic north-easterly
HFA	Scandinavian high, ridge over central Europe
HFZ	Scandinavian high, trough over central Europe
HNFA	Scandinavian-Iceland high, ridge over central Europe
HNFZ	Scandinavian-Iceland high, trough over central Europe
SEA	Anticyclonic south-easterly
SEZ	Cyclonic south-easterly
SA	Anticyclonic southerly
SZ	Cyclonic southerly
TB	Low over the British Isles
TRW	Trough over western Europe

Appendix B: Statistical background

The works collected here present some results for a deeper understanding of the 'Statistical background' section in the main text of the paper. This is not a presentation of new results: it is intended to collate the background information that is typically found in undergraduate statistics courses in an accessible way. Much of the information presented in this section can be found in DeGroot and Schervish (2012).

Estimators and estimates

Consider a general random variable X taking values in $\mathcal{X}$ with probability density function (PDF) $f(\cdot;\theta) : \mathcal{X} \to \mathbb{R}$, characterised by some parameter $\theta \in \Theta \subseteq \mathbb{R}$. We write $X \sim f(\cdot;\theta)$, or simply $X \sim f(\theta)$, to show that the distribution of X is described by f for some given value of the parameter $\theta \in \Theta$.

If we consider n random samples from the distribution of X for some unknown value $\theta^* \in \Theta$, we write $\mathbf{X} \sim f_n(\theta^*)$ with $\mathbf{X} := (X_1, \ldots, X_n)$, where f_n describes the joint distribution of $\mathbf{X}$. In particular, if $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.), we denote this by $\mathbf{X} \overset{\text{iid}}{\sim} f_n(\theta^*)$, where the joint density of $\mathbf{X}$ is the product of the densities of the X_i:

$$f_n(\mathbf{x};\theta^*) = \prod_{i=1}^{n} f(x_i;\theta^*). \tag{B1}$$

Definition 1 (Estimate and estimator). Suppose that $\mathbf{X} \sim f_n(\theta^*)$ for some unknown $\theta^* \in \Theta$ and let $\mathbf{x}$ be a realisation of $\mathbf{X}$. For some function $\hat{\theta} : \mathcal{X}^n \to \Theta$ of $\mathbf{x}$ that is intended to provide an approximation of θ^*, $\hat{\theta}(\mathbf{x})$ is an *estimate* of θ^*. Then $\hat{\theta}$, or $\hat{\theta}(\mathbf{X})$, is referred to as an *estimator* of θ^*.

We emphasise that the estimator is a random variable and the estimate a realisation of this random variable, but the terms may be used interchangeably in the main text of the paper.

Definition 2 (Unbiased). $\hat{\theta}$ is an *unbiased* estimator of θ if, for all values $\theta \in \Theta$, its expected value is the true value θ: $\mathbb{E}\left[\hat{\theta}(\mathbf{X})\right] = \theta$.

An unbiased estimator $\hat{\theta}$ thus provides us with unbiased estimates of θ given data $x_1, \ldots, x_n$. The performance of an estimator, in terms of the discrepancy between the estimated value and its true value, can be measured via a loss function such as the mean squared error.

Definition 3 (Mean squared error). The *mean squared error* (MSE) of an estimator $\hat{\theta}$, given $\mathbf{X} \sim f_n(\theta^*)$ for some $\theta^* \in \Theta$, is a loss function defined by

$$\text{MSE}\left(\hat{\theta}(\mathbf{X})\right) := \mathbb{E}\left[\left(\hat{\theta}(\mathbf{X}) - \theta^*\right)^2\right], \tag{B2}$$

the mean of the expected squared difference between the estimator and the true value θ^*.

Equation B2 can be decomposed into the variance and squared bias of the estimator, referred to as the *bias-variance decomposition*:

$$\text{MSE}\left(\hat{\theta}(\mathbf{X})\right) = \text{Var}\left(\hat{\theta}(\mathbf{X})\right) + \left(\text{bias}\left(\hat{\theta}(\mathbf{X})\right)\right)^2, \tag{B3}$$

where we define the bias as the expected difference between the value of the estimator and the true value of the parameter:

$$\text{bias}\left(\hat{\theta}(\mathbf{X})\right) := \mathbb{E}\left[\hat{\theta}(\mathbf{X}) - \theta^*\right] = \mathbb{E}\left[\hat{\theta}(\mathbf{X})\right] - \theta^*. \tag{B4}$$

Remark 1 Noting that $\text{bias}\left(\hat{\theta}(\mathbf{X})\right) = 0$ when the estimator is unbiased, the bias-variance decomposition illustrates that the MSE of an unbiased estimator is its variance.

Asymptotic behaviour of estimators

Let us denote by $\{X_n\}$ a sequence $X_1, X_2, \ldots$ of random variables taking values in $\mathcal{X}$ with the same distribution, indexed by n. In this section, we introduce the notion of probabilistic convergence of random variables and their distributions. This is key to understanding the reasoning behind why we can construct approximate confidence intervals for sufficiently large n using a standard normal distribution.

Definition 4 (Convergence in probability). $\{X_n\}$ is said to *converge in probability* to the random variable X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0, \tag{B5}$$

which we write as $X_n \to_{\mathbb{P}} X$.

Definition 5 (Convergence in distribution). $\{X_n\}$ is said to *converge in distribution* to a random variable X if, for all x where the cumulative distribution function (CDF) $F_X(x) := \mathbb{P}(X \le x)$ is continuous,

$$\lim_{n\to\infty} \mathbb{P}(X_n \le x) = \mathbb{P}(X \le x)\,. \tag{B6}$$

Then, the distribution of X is referred to as the *asymptotic distribution* of $\{X_n\}$, and we write this as $X_n \to_{\mathcal{D}} X$.

We present the following theorems in relation to the asymptotic distributions of sequences of random variables.

Theorem 1 (Weak Law of Large Numbers) *If $\{X_n\}$ are i.i.d. random variables with common expectation $\mathbb{E}[X_1] = \mu < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{\mathbb{P}} \mu. \tag{B7}$$

Theorem 2 (Central Limit Theorem) *If $X_1, \ldots, X_n$ are i.i.d. random variables such that $\mathbb{E}[X_i] = \mu < \infty$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$ for all $i \in \{1, \ldots, n\}$, then*

$$\frac{\sqrt{n}(S_n - \mu)}{\sigma} \to_{\mathcal{D}} Z \sim \mathcal{N}(0, 1), \tag{B8}$$

where $S_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.

Remark 2 We may write $\sqrt{n}(S_n - \mu)/\sigma \to_{\mathcal{D}} Z \sim \mathcal{N}(0, 1)$ as $\sqrt{n}(S_n - \mu)/\sigma \to_{\mathcal{D}} \mathcal{N}(0, 1)$, a slight abuse of notation, for simplicity.

Theorem 3 *Convergence in probability implies convergence in distribution.*

Theorem 4 (Slutsky's theorem) *If $Y_n \to_{\mathcal{D}} Y$ and $Z_n \to_{\mathcal{D}} c$, where $c \in \mathbb{R}$ is a constant, then*

1. $Y_n + Z_n \to_{\mathcal{D}} Y + c$.
2. $Y_n Z_n \to_{\mathcal{D}} Yc$.
3. $Y_n / Z_n \to_{\mathcal{D}} Y/c$, *provided* $c \neq 0$.

Theorem 5 (Continuous Mapping Theorem) *Let $g : \mathcal{X} \to \mathcal{G}$ be continuous. Then,*

1. *If $Z_n \to_{\mathbb{P}} Z$, then $g(Z_n) \to_{\mathbb{P}} g(Z)$.*
2. *If $Z_n \to_{\mathcal{D}} Z$, then $g(Z_n) \to_{\mathcal{D}} g(Z)$.*

Furthermore, we define the notions of consistency for a sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots$, which we denote by $\{\hat{\theta}_n\}$, and of asymptotic normality. This is essential for the construction of asymptotically exact confidence intervals for θ.

Definition 6 (Consistent). A sequence of estimators $\{\hat{\theta}_n\}$ is *consistent* if, for all $\theta \in \Theta$ with $\mathbf{X} \sim f_n(\theta)$,

$$\hat{\theta}_n(\mathbf{X}) \to_{\mathbb{P}} \theta. \tag{B9}$$

Definition 7 (Asymptotic normality). Let $\{\hat{\theta}_n\}$ denote a consistent sequence of estimators for some parameter $\theta \in \Theta$. $\{\hat{\theta}_n\}$ is *asymptotically normal* if, for some $\sigma^2 > 0$,

$$\frac{\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta)}{\sigma} \to_{\mathcal{D}} \mathcal{N}(0, 1). \tag{B10}$$

The delta method illustrates that an asymptotically normal estimator remains asymptotically normal under transformation by a continuously differentiable function (Doob, 1935).

Theorem 6 (Delta method) *Suppose $\{\hat{\theta}_n\}$ is a sequence of consistent and asymptotically normal estimators and $g : \Theta \to \mathcal{G}$ is a continuously differentiable function whose first derivative $g'(\theta)$ is continuous and non-zero for all $\theta \in \Theta$. Then,*

$$\frac{\sqrt{n}\left(g(\hat{\theta}_n(\mathbf{X})) - g(\theta)\right)}{\sigma g'(\theta)} \to_{\mathcal{D}} \mathcal{N}(0, 1), \tag{B11}$$

i.e. $\{g(\hat{\theta}_n(\mathbf{X})\}$ is asymptotically normal.

Asymptotically exact confidence intervals

Definition 8 (Confidence interval). Let $\mathbf{X} \sim f_n(\theta)$ for some $\theta \in \Theta$ and let $L : \mathcal{X}^n \to \Theta$ and $U : \mathcal{X}^n \to \Theta$ be functions satisfying $L(\mathbf{x}) < U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^n$. For some $\alpha \in [0, 1]$, we say that a random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is a $1-\alpha$ *confidence interval* if, for all $\theta \in \Theta$,

$$\begin{aligned}\mathbb{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]) &= \mathbb{P}(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) \\ &\geq 1 - \alpha.\end{aligned}$$

We refer to $\mathbb{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ as the *coverage* of $[L(\mathbf{X}), U(\mathbf{X})]$.

Remark 3 The above definition of a confidence interval is a probability statement about the joint distribution of the random variables $L(\mathbf{X})$ and $U(\mathbf{X})$, given a particular value of θ. Once the values of $X_1, \ldots, X_n$ are observed to be $x_1, \ldots, x_n$, we compute the values of $L(\mathbf{X})$ and $U(\mathbf{X})$ to obtain an *observed* confidence interval $[L_n(\mathbf{x}), U_n(\mathbf{x})]$.

The following theorem tells us how to compute an observed confidence interval such that the coverage of the interval approaches $1 - \alpha$ as the sample size increases.

Theorem 7 (Asymptotically exact confidence intervals) *Suppose $\{\hat{\theta}_n\}$ is a consistent sequence of estimators of θ that is asymptotically normal, and also that $\{\hat{\sigma}_n^2\}$ is a consistent sequence of estimators of σ^2. Then, for all $\alpha \in (0, 1)$, $[L_n(\mathbf{x}), U_n(\mathbf{x})]$ is an asymptotically exact $1 - \alpha$ confidence interval for θ, where*

$$L_n(\mathbf{x}) = \hat{\theta}_n(\mathbf{x}) - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2(\mathbf{x})}{n}}, \tag{B12}$$

$$U_n(\mathbf{x}) = \hat{\theta}_n(\mathbf{x}) + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2(\mathbf{x})}{n}}. \tag{B13}$$

The confidence interval is only asymptotically exact, so in practice, the coverage of the confidence interval will be different from $1-\alpha$. However, as n increases, the coverage will tend towards $1-\alpha$.

Suppose we parameterise the problem in terms of $\tau := g(\theta)$, where g is a bijective and continuously differentiable function. In order to find a $1-\alpha$ confidence interval for τ, we might first consider a direct transformation of the confidence interval $[L(\mathbf{x}), U(\mathbf{x})]$ for θ^* through noting that

$$\begin{aligned}&\left\{\mathbf{x}\in\mathcal{X}^n : L(\mathbf{x})\le\theta^*\le U(\mathbf{x})\right\}\\&=\left\{\mathbf{x}\in\mathcal{X}^n : g(L(\mathbf{x}))\le\tau^*=g(\theta^*)\le g(U(\mathbf{x}))\right\},\end{aligned}\tag{B14}$$

if g is an increasing function. We find that, as $n\to\infty$,

$$\begin{aligned}&\mathbb{P}\big(L(\mathbf{x})\le\theta^*\le U(\mathbf{x})\big)\\&=\mathbb{P}\big(g(L(\mathbf{x}))\le\tau^*=g(\theta^*)\le g(U(\mathbf{x}))\big)\to 1-\alpha,\end{aligned}\tag{B15}$$

so $[g(L(\mathbf{x})), g(U(\mathbf{x}))]$ is an asymptotically exact $1-\alpha$ confidence interval for τ^*. For decreasing g, we simply reverse the directions of the inequalities in the above. Note that this approach does not necessarily obtain a confidence interval that is centred at $\hat{\tau}_n(\mathbf{x})$.

Alternatively, we note from Theorem 6 that if $\{\hat{\theta}_n\}$ is a consistent sequence of estimators that is asymptotically normal with mean θ and variance σ^2, then $\{\hat{\tau}_n\}$ is also a consistent and asymptotically normal sequence of estimators with mean $g(\theta)=\tau$ and variance $\sigma^2(g'(\theta))^2$. Then, we can simply reapply Theorem 7 to $\{\hat{\tau}_n\}$, yielding an asymptotically exact $1-\alpha$ confidence interval $[\bar{L}(\mathbf{x}), \bar{U}(\mathbf{x})]$ for τ:

$$\bar{L}(\mathbf{x}) := \hat{\tau}_n(\mathbf{x}) - z_{\alpha/2}\sqrt{\frac{\sigma_n^2(\mathbf{x})(g'(\theta_n(\mathbf{x}))^2}{n}},\tag{B16}$$

$$\bar{U}(\mathbf{x}) := \hat{\tau}_n(\mathbf{x}) + z_{\alpha/2}\sqrt{\frac{\sigma_n^2(\mathbf{x})(g'(\theta_n(\mathbf{x}))^2}{n}}.\tag{B17}$$

Example 1 For a random variable X with distribution function f, a common exercise is to estimate the expected value of X, $\mu=\mathbb{E}[X]$. We sample n times from the distribution of $X_1,\ldots,X_n$ and estimate μ by $\hat{\mu}_n=\frac{1}{n}\sum_{i=1}^n X_i$. A consistent estimator for $\sigma^2=\mathrm{Var}(X)$ is then $\hat{\sigma}_n^2=\frac{1}{n}\sum_{i=1}^n(X_i-\hat{\mu}_n)^2$, so that a $1-\alpha$ confidence interval for μ is given by

$$\hat{\mu}_n \pm z_{\alpha/2}\sqrt{\frac{\sum_{i=1}^n(X_i-\hat{\mu}_n)^2}{n}}.\tag{B18}$$

Suppose we now wish to find an estimate for $\log\mu$ and a corresponding confidence interval. From Theorem 5, we know that $\log\hat{\mu}_n\to_{\mathbb{P}}\log\mu$, so that $\log\hat{\mu}_n$ is consistent. Furthermore, since $\log\mu$ is continuously differentiable with non-zero first derivative $1/\mu$, we can apply Theorem 6 to show that the sequence of estimators is asymptotically normal and use Eqs. B16 and B17 to obtain a confidence interval for $\log\mu$:

$$\log\hat{\mu}_n \pm z_{\alpha/2}\sqrt{\frac{\sum_{i=1}^n(X_i-\hat{\mu}_n)^2}{n\hat{\mu}_n}}.\tag{B19}$$

Estimation of Bernoulli probabilities

For the Bernoulli(p) random variable X, which we define in the main text to represent exceedance for some threshold c, we aim to estimate the value of the exceedance probability $p\in(0,1)$ given some realised observations $\mathbf{x}$ drawn from the distribution of $\mathbf{X}\overset{\text{iid}}{\sim} f_n(p)$, where

$$f(x;p)=p^x(1-p)^x\tag{B20}$$

for $x\in\{0,1\}$, and hence

$$f_n(\mathbf{x};p)=\prod_{i=1}^n f(x;p)=p^{x_i}(1-p)^{1-x_i}\tag{B21}$$

for $\mathbf{x}=(x_1,\ldots,x_n)\in\mathcal{X}^n$. X has expectation $\mathbb{E}[X]=p$ and variance $\sigma^2=\mathrm{Var}(X)=p(1-p)$, and we define the simple estimator of p as

$$p^n_{\text{simple}} := \frac{1}{n}\sum_{i=1}^n X_i,\tag{B22}$$

which is unbiased,

$$\mathbb{E}\left[p^n_{\text{simple}}\right]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[X_i]=\frac{1}{n}np=p,\tag{B23}$$

with variance given by

$$\begin{aligned}\mathrm{Var}\left(p^n_{\text{simple}}\right)&=\frac{1}{n^2}\sum_{i=1}^n\mathrm{Var}(X_i)=\frac{1}{n^2}np(1-p)\\&=\frac{p(1-p)}{n}.\end{aligned}\tag{B24}$$

Since $p^n_{\text{simple}}=\frac{1}{n}\sum_{i=1}^n X_i$, where the X_i are i.i.d. with $\mathbb{E}[X_i]=p$ for all $i=1,\ldots,n$, it follows from Theorem 1 that $p^n_{\text{simple}}\to_{\mathbb{P}} p$, i.e. it is a consistent estimator of p. Furthermore, by Theorem 2, p^n_{simple} is asymptotically normal.

Using Theorem 7, we obtain the asymptotically exact confidence intervals Eq. 5 for p:

$$p^n_{\text{simple}} \pm z_{\alpha/2}\sqrt{\frac{p^n_{\text{simple}}(1-p^n_{\text{simple}})}{n}}.\tag{B25}$$

Remark 4 To obtain the confidence intervals for $\log p$, we define the reparameterisation $\tau := g(p) = \log p$. We use the result in Example 1 to obtain the asymptotically exact confidence intervals Eq. 6 for $\log p$:

$$\log\left(p^n_{\text{simple}}\right) \pm z_{\alpha/2}\sqrt{\frac{1-p_{\text{simple}}}{np_{\text{simple}}}}. \tag{B26}$$

Appendix C: Results related to stratified sampling

Variance decomposition into within- and between-stratum components

We show that Eq. 17 holds, i.e. the variance of a Bernoulli(p) random variable X can be decomposed into within- and between-stratum components. Let Z be the random variable representing the start date of the wind field data drawn uniformly at random from $\mathcal{Z}$, where $\mathcal{Z}$ is partitioned into J strata $\mathcal{Z}_1, \ldots, \mathcal{Z}_J$. For ease of notation, we represent the event $\{Z \in \mathcal{Z}_j\}$ by $\{Y = j\}$, where Y is a random variable drawn from $\{1, \ldots, J\}$ with probability

$$\mathbb{P}(Y = j) = \mathbb{P}\big(Z \in \mathcal{Z}_j\big) = w_j, \tag{C1}$$

for all $j \in \{1, \ldots, J\}$, and $\sum_{j=1}^J w_j = 1$. Noting that

$$\mathbb{E}[X|Y = j] = p_j, \tag{C2}$$
$$\text{Var}(X|Y = j) = p_j(1 - p_j), \tag{C3}$$

for all $j \in \{1, \ldots, J\}$, we obtain

$$\begin{aligned}\mathbb{E}[\text{Var}(X|Y)] &= \sum_{j=1}^J \mathbb{P}(Y = j)\,\text{Var}(X|Y = j)\\ &= \sum_{j=1}^J w_j p_j(1 - p_j);\end{aligned} \tag{C4}$$

$$\begin{aligned}\text{Var}(\mathbb{E}[X|Y]) &= \mathbb{E}\Big[(\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Y]])^2\Big]\\ &= \mathbb{E}\Big[(\mathbb{E}[X|Y] - \mathbb{E}[X])^2\Big]\\ &= \sum_{j=1}^J \mathbb{P}(Y = j)\,(\mathbb{E}[X|Y = j] - \mathbb{E}[X])^2\\ &= \sum_{j=1}^J w_j(p_j - p)^2.\end{aligned} \tag{C5}$$

Then, by the law of total variance,

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y]) \tag{C6}$$
$$= \sum_{j=1}^J w_j p_j(1 - p_j) + \sum_{j=1}^J w_j(p_j - p)^2. \tag{C7}$$

Unbiasedness and variance of the stratified sampling estimator

For the stratified sampling estimator p^n_{strat} defined in Eq. 14, we show that it is an unbiased estimator of p. Noting that the samples $X_{j,1}, \ldots, X_{j,n_j}$ are independent and identically distributed so that $\mathbb{E}\big[X_{j,i}\big] = \mathbb{E}\big[X_{j,1}\big]$ for all $j \in \{1, \ldots, J\}$, we have

$$\begin{aligned}\mathbb{E}\big[p^n_{\text{strat}}\big] &= \mathbb{E}\left[\sum_{j=1}^J w_j \frac{1}{n_j}\sum_{i=1}^{n_j} X_{j,i}\right]\\ &= \sum_{j=1}^J w_j p_j = p,\end{aligned} \tag{C8}$$

so it is unbiased for all $p \in (0, 1)$. Furthermore, the variance of p^n_{strat} is given by

$$\begin{aligned}\text{Var}\left(p^n_{\text{strat}}\right) &= \sum_{j=1}^J w_j^2 \text{Var}\left(p^n_j\right)\\ &= \sum_{j=1}^J \frac{w_j^2}{n_j} p_j(1 - p_j).\end{aligned} \tag{C9}$$

In particular, if $n_j = nw_j$ for all $j \in \{1, \ldots, J\}$, as in proportional stratum allocation, we find that the variance of the estimator is guaranteed to be less than or equal to that of p^n_{simple}:

$$\begin{aligned}\text{Var}\left(p^n_{\text{strat}}\right) &= \sum_{j=1}^J \frac{w_j^2}{nw_j} p_j(1 - p_j)\\ &= \frac{1}{n}\sum_{j=1}^J w_j p_j(1 - p_j)\\ &\le \frac{1}{n}\sum_{j=1}^J w_j\left(p_j(1 - p_j) + (p_j - p)^2\right)\\ &= \frac{1}{n}\text{Var}(X) = \text{Var}\left(p^n_{\text{simple}}\right),\end{aligned} \tag{C10}$$

where the final line follows from Eq. 17.

Asymptotic normality of the stratified sampling estimator

We show that the stratified sampling estimator is asymptotically normal. Let $q_j := \lim_{n\to\infty} n_j/n$.

$$\begin{aligned}\sqrt{n}\left(p^n_{\text{strat}} - p\right) &= \sqrt{n}\left(\sum_{j=1}^{J}\frac{w_j}{n_j}\sum_{i=1}^{n_j}X_{j,i} - \sum_{j=1}^{J}w_j p_j\right)\\ &= \sqrt{n}\left(\sum_{j=1}^{J}w_j\left(\frac{1}{n_j}\sum_{j=1}^{J}X_{j,i} - p_j\right)\right)\\ &= \sqrt{n}\left(\sum_{j=1}^{J}\frac{w_j}{\sqrt{n_j}}\left(\frac{1}{\sqrt{n_j}}S_j\right)\right)\\ &= \sum_{j=1}^{J}\frac{w_j}{\sqrt{n_j/n}}\left(\frac{1}{\sqrt{n_j}}S_j\right), \end{aligned} \tag{C11}$$

where we denote $S_j := \sum_{i=1}^{n_j} X_{j,i} - p_j$. By Theorem 2, $S_j/\sqrt{n_j} \to_{\mathcal{D}} \mathcal{N}(0, \sigma_j^2)$. It follows that

$$\frac{w_j}{\sqrt{n_j/n}}\frac{1}{n_j}S_j \to_{\mathcal{D}} \mathcal{N}\left(0, \frac{w_j^2}{q_j}\sigma_j^2\right). \tag{C12}$$

Then, since the J terms in the summation in Eq. C11 are independent, we obtain

$$\begin{aligned}\sqrt{n}\left(p^n_{\text{strat}} - p\right) &= \sum_{j=1}^{J}\frac{w_j}{\sqrt{n_j/n}}\frac{1}{n_j}S_j\\ &\to_{\mathcal{D}} \mathcal{N}\left(0, \sum_{j=1}^{J}\frac{w_j^2}{q_j}\sigma_j^2\right). \end{aligned} \tag{C13}$$

That is, the stratified sampling estimator is asymptotically normal with mean p and variance $\frac{1}{n}\sum_{j=1}^{J} w_j^2\sigma_j^2/q_j$, where $q_j = \lim_{n\to\infty} n_j/n$.

Optimal stratum allocation

It is possible to define an optimum stratification which will always reduce the variance of the estimate, where the number of samples is allocated proportionally to both the weights w_j and the stratum standard deviations σ_j, where $\sigma_j^2 := p_j(1-p_j)$ for Bernoulli probabilities p_j (see Etore and Jourdain 2010). The variance of the optimal stratified estimator is guaranteed to be less than that of proportional allocation, but requires knowledge of the stratum variances. Since these are usually unknown, approximately optimal stratified estimators can be constructed using accurate approximations of the stratum variances. In the setting of exceedance probability estimation, stratum variances are linked to the exceedance threshold of interest; if a range of thresholds are considered, optimal allocation using any one threshold may not necessarily provide better results overall. Therefore, in this paper, we did not pursue optimal stratification, but here, we note its benefit in variance reduction if sufficiently accurate approximations of stratum variances are available.

Post-stratification

If our samples have already been drawn according to random sampling, it is possible to post-hoc reduce the variance of our probability estimates through a variant of stratified sampling called *post-stratification* (Jagers et al., 1985). The method has applications in areas such as political polling, where survey responses are biased and require debiasing to be representative of the entire population (Jagers, 1986). An example application of post-stratification to the use of weather regimes in volcanic ash hazard assessment would be if the frequency of the weather regimes changed in the future (without a change in the variation within patterns), or perhaps if there were known seasonal changes in the frequencies. Post-stratification would allow re-weighting of the patterns without needing to resample individual wind fields from their distribution.

In particular, using the notation described in the 'Statistical background' section of the main paper, if the simulation start dates $Z_1, \ldots, Z_n$ have been drawn independently with probabilities $q_1, \ldots, q_J$, we may allocate each sample to the appropriate stratum according to their classification to obtain stratum sizes

$$n_j = \left|\left\{i : Z_i \in \mathcal{Z}_j,\ i = 1, \ldots, n\right\}\right| \tag{C14}$$

for $j \in \{1, \ldots, J\}$, which we note are now random. We then denote by $X_{j,1}, \ldots, X_{j,n_j}$ the corresponding Bernoulli(p_j) random variables representing exceedance for some threshold. Stratum sample means can then be computed by Eq. 13 and the stratified sampling estimator of p by Eq. 14, which we refer to as the *post-stratified estimator*.

The caveat of this method, however, is that the post-stratified estimate cannot be computed if $n_j = 0$ for some $j \in \{1, \ldots, J\}$ (Jagers et al., 1985). As the n_j are random, the probability of this event is

$$\mathbb{P}\left(\min_j n_j = 0\right) = \sum_{j=1}^{J}(1-q_j)^n, \tag{C15}$$

which decreases exponentially quickly in n with a rate depending on the smallest weight and hence becomes very small as n increases.

In the following section, we show that the post-stratified estimator is unbiased whenever the stratum sizes are non-zero. Moreover, the estimator is asymptotically normal which allows us to construct asymptotically exact confidence intervals.

Unbiasedness and asymptotic normality of the post-stratified estimator

We view post-stratification as a situation in which the stratum sizes $n_1, \ldots, n_J$ are given positive integers. Given $n_j > 0$ for each $j \in \{1, \ldots, J\}$, $X_{j,i} \sim \text{Bernoulli}(p_j)$ for each $i \in \{1, \ldots, n_j\}$. Then, conditional on the n_j being positive, we are in the standard stratified sampling setting, and the post-stratified estimator is unbiased with the same asymptotic distribution as in the previous section, provided that $\lim_{n\to\infty} n_j/n = q_j$.

If $n_j = 0$ for some $j \in \{1, \ldots, J\}$, then the post-stratified estimator is not defined. It is possible to define the estimator in some other way in this case; however, this would introduce bias. Therefore, the post-stratified estimator is unbiased whenever $\min_{j\in\{j,\ldots,J\}} n_j > 0$.

An alternative approach is to treat the $n_1, \ldots, n_j$ as realisations of random variables $N_1, \ldots, N_J$ as in Jagers et al. (1985). Let $Z_1, \ldots, Z_n$ be independent random variables drawn from $\mathcal{Z}$ such that $q_j := \mathbb{P}(Z \in \mathcal{Z}_j)$ for all $j \in \{1, \ldots, J\}$, where $\sum_{j=1}^{J} q_j = 1$. Then,

$$N_j = \sum_{i=1}^{n} \mathbb{1}\{Z_i \in \mathcal{Z}_j\}, \tag{C16}$$

and we see that $(N_1, \ldots, N_J) \sim \text{Multinomial}(n, (q_1, \ldots, q_J))$. Letting $f(Z_i) = \mathbb{1}\{\psi \circ \phi(Z_i) \geq c\} \sim \text{Bernoulli}(p)$ represent the exceedance event, we write the post-stratified estimator as

$$p^n_{\text{post}} := \sum_{j=1}^{J} w_j \frac{1}{N_j} \sum_{i=1}^{n} f(Z_i)\, \mathbb{1}\{Z_i \in \mathcal{Z}_j\}. \tag{C17}$$

Then,

$$\begin{aligned}
&\sqrt{n}\left(p^n_{\text{post}} - p\right) \\
&= \sqrt{n}\left(\sum_{j=1}^{J} w_j \frac{1}{N_j} \sum_{i=1}^{n} \mathbb{1}\{Z_i \in \mathcal{Z}_j\}\left(f(Z_i) - p_j\right)\right) \\
&= \sum_{j=1}^{J} \frac{n w_j}{N_j} \frac{1}{\sqrt{n}} S_j,
\end{aligned} \tag{C18}$$

where we define

$$S_j = \sum_{i=1}^{n} \mathbb{1}\{Z_i \in \mathcal{Z}_j\}\left(f(Z_i) - p_j\right) =: \sum_{i=1}^{n} \varphi_j(Z_i). \tag{C19}$$

Noting that

$$\begin{aligned}
\varphi_j(Z_i) &= \begin{cases} -p_j & \text{if } Z_i \in \mathcal{Z}_j \text{ and } X_i = 0, \\ 1 - p_j & \text{if } Z_i \in \mathcal{Z}_j \text{ and } X_i = 1, \\ 0 & \text{if } Z_i \notin \mathcal{Z}_j, \end{cases} \\
&= \begin{cases} -p_j & \text{w.p. } q_j(1 - p_j), \\ 1 - p_j & \text{w.p. } q_j p_j, \\ 0 & \text{w.p. } 1 - q_j, \end{cases}
\end{aligned} \tag{C20}$$

we find that

$$\mathbb{E}\left[\varphi_j(Z_i)\right] = (1 - p_j) q_j p_j - p_j q_j (1 - p_j) = 0 \tag{C21}$$

and

$$\begin{aligned}
\mathbb{E}\left[\left(\varphi_j(Z_i)\right)^2\right] &= (1 - p_j)^2 q_j p_j + p_j^2 q_j (1 - p_j) \\
&= q_j p_j (1 - p_j) = q_j \sigma_j^2,
\end{aligned} \tag{C22}$$

giving

$$\begin{aligned}
\text{Var}\left(\varphi_j(Z_i)\right) &= \mathbb{E}\left[\left(\varphi_j(Z_i)\right)^2\right] - \left(\mathbb{E}\left[\varphi_j(Z_i)\right]\right)^2 \\
&= \mathbb{E}\left[\left(\varphi_j(Z_i)\right)^2\right] = q_j \sigma_j^2.
\end{aligned} \tag{C23}$$

It follows from Theorem 2 that $\frac{1}{\sqrt{n}} S_j \to_{\mathcal{D}} \bar{S}_j \sim \mathcal{N}\left(0, q_j \sigma_j^2\right)$, for all $j \in \{1, \ldots, J\}$. Noting that $\mathbb{E}[\varphi_j(Z_i)\varphi_k(Z_i)] = 0$ for all $j \neq k$, we show that the $\varphi_j(Z_i)$ are uncorrelated for all $i \in \{1, \ldots, n\}$:

$$\begin{aligned}
&\text{Cov}\left(\varphi_j(Z_i), \varphi_j(Z_i)\right) \\
&= \mathbb{E}\left[\left(\varphi_j(Z_i) - \mathbb{E}\left[\varphi_j(Z_i)\right]\right)\left(\varphi_k(Z_i) - \mathbb{E}[\varphi_k(Z_i)]\right)\right] \\
&= \mathbb{E}\left[\varphi_j(Z_i)\varphi_k(Z_i)\right] = 0,
\end{aligned} \tag{C24}$$

which follows from Eq. C21. Using this result, we arrive at

$$\text{Cov}\left(\frac{1}{\sqrt{n}} S_j, \frac{1}{\sqrt{n}} S_k\right) = \mathbb{E}\left[S_j S_k\right] = 0, \tag{C25}$$

for $j \neq k$. Then, $\frac{1}{\sqrt{n}} S_1, \ldots, \frac{1}{\sqrt{n}} S_J$ are uncorrelated and converge in distribution to $\bar{S}_1, \ldots, \bar{S}_J$ which are normal and uncorrelated, and hence independent. Furthermore, by Theorem 1, $n w_j / N_j \to_{\mathbb{P}} w_j / q_j$. It then follows from Theorem 4 that

$$\frac{n w_j}{N_j} \frac{1}{\sqrt{n}} S_j \to_{\mathcal{D}} \mathcal{N}\left(0, \frac{w_j^2}{q_j} \sigma_j^2\right). \tag{C26}$$

Since there is joint convergence in distribution of $\frac{1}{\sqrt{n}}S_1, \ldots, \frac{1}{\sqrt{n}}S_J$ to independent normal random variables, and in $nw_1/N_1, \ldots, nw_J/N_J$ to constants, we can apply Theorem 5 to show that

$$\sqrt{n}\left(p^n_{\text{post}} - p\right) = \sum_{j=1}^{J} \frac{nw_j}{N_j}\frac{1}{\sqrt{n}}S_j \rightarrow_{\mathcal{D}} \mathcal{N}\left(0, \sum_{j=1}^{J}\frac{w_j^2}{q_j}\sigma_j^2\right), \quad \text{(C27)}$$

which is the same as Eq. C13. That is, the post-stratified estimator has the same asymptotic distribution as the original stratified sampling estimator despite the samples $Z_1, \ldots, Z_n$ being drawn from some other distribution.

Supplementary information

An R package for carrying out the statistical methods discussed is available on GitHub (https://github.com/shannon-wms/stratsampling).

Acknowledgements This research was supported by funding to JP and SJ from the UK Natural Environmental Research Council Environmental Risks to Infrastructure Impact's project NE/M008878/1 and by EDF subcontract 56794 to University of Bristol, and to SW and AL from the UK Engineering and Physical Sciences Research Council (EP/S023569/1 and EP/R034710/1, respectively).
We gratefully acknowledge useful discussions about this work with Willy Aspinall, Henry Odbert, David Parker, Pietro Bernardara, Katie Fish, Keval Nakeshree, and Paul Tucker.

Author Contributions JP and SW are joint first authors on this publication. JP conceived the overall study and led the drafting of the manuscript with SW and SJ. SW developed and conducted the statistical analyses with support from AL, drafted the statistical content of the paper, and drafted the results figures. SJ and JP conducted the dispersion modelling and developed the concept of using weather regimes. All authors contributed to writing and reviewing the manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agresti A, Coull BA (1998) Approximate is better than "exact" for interval estimation of binomial proportions. Am Sta 52(2):119–126. https://doi.org/10.2307/2685469

Ata MY (2007) A convergence criterion for the Monte Carlo estimates. Simul Model Pract Theory 15(3):237–246. https://doi.org/10.1016/j.simpat.2006.12.002

Barnston AG, Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Monthly weather review 115(6):1083–1126. https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2

Baur F, Hess P, Nagel H (1944) Kalender der Grosswetterlagen Europas 1881–1939. Bad Homburg 35

Biass S, Scaini C, Bonadonna C et al (2014) A multi-scale risk assessment for tephra fallout and airborne concentration from multiple Icelandic volcanoes-Part 1: Hazard assessment. Hazards Earth Syst Sci 14:2265–2287. https://doi.org/10.5194/nhess-14-2265-2014

Biass S, Bonadonna C, Connor L et al (2016) TephraProb: a Matlab package for probabilistic hazard assessments of tephra fallout. J Appl Volcanol 5(1):1–16. https://doi.org/10.1186/s13617-016-0050-5

Bonadonna C (2006) Probabilistic modelling of tephra dispersion. Statistics in Volcanology Special Publications of IAVCEI (Geological Society, London) 1:243–259. https://doi.org/10.1144/iavcei001.19

Bonadonna C, Connor CB, Houghton BF et al (2005) Probabilistic modeling of tephra dispersal: hazard assessment of a multiphase rhyolitic eruption at Tarawera. New Zealand. J Geophys Res Solid 110(B3) https://doi.org/10.1029/2003JB002896

Brown LD, Cai TT, Dasgupta A (2001) Interval estimation for a binomial proportion. Stat Sci 16(2):101–117. https://doi.org/10.1214/ss/1009213286

Capponi A, Harvey NJ, Dacre HF et al (2022) Refining an ensemble of volcanic ash forecasts using satellite retrievals: Raikoke 2019. Atmos Chem Phys 22(9):6115–6134. https://doi.org/10.5194/acp-22-6115-2022

Cochran WG (1977) Sampling techniques (Third Edition). John Wiley and Sons

Connor CB, Hill BE, Winfrey B et al (2001) Estimation of volcanic hazards from tephra fallout. Nat Hazards Rev 2(1):33–42. https://doi.org/10.1061/(ASCE)1527-6988(2001)2:1(33)

Crosweller HS, Arora B, Brown SK et al (2012) Global database on large magnitude explosive volcanic eruptions (LaMEVE). J Appl Volcanol 1(1):1–13. https://doi.org/10.1186/2191-5040-1-4

Junior Da Silva Fonseca JG, Oozeki T, Ohtake H, et al (2015) Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods. Prog Photovolt 23(10):1203–1218. https://doi.org/10.1002/pip.2528

D'Amato V, Haberman S, Russolillo M (2012) The stratified sampling bootstrap for measuring the uncertainty in mortality forecasts. Methodol Comput Appl 14(1):135–148. https://doi.org/10.1007/S11009-011-9225-Z

DeGroot MH, Schervish MJ (2012) Probability and statistics. Pearson Education

Doob JL (1935) The limiting distributions of certain statistics. Ann Math Stat 6(3):160–169. https://doi.org/10.1214/aoms/1177732594

DWD (2015) Large-scale weather forecasting (GWL). Tech. rep., Deutscher Wetterdienst, https://www.dwd.de/DE/forschung/wettervorhersage/met_fachverfahren/nwv_anschlussverfahren/grosswetterlagen_klassifizierung.html, Accessed on 19 Dec 2022

Étoré P, Jourdain B (2010) Adaptive optimal allocation in stratified sampling methods. Methodol Comput Appl Probab 12(3):335–360. https://doi.org/10.1007/s11009-008-9108-0. arXiv:0711.4514

Folch A, Costa A, Macedonio G (2009) FALL3D: a computational model for transport and deposition of volcanic ash. Comput Geosci 35(6):1334–1342. https://doi.org/10.1016/j.cageo.2008.08.008

Folch A, Costa A, Basart S (2012) Validation of the FALL3D ash dispersion model using observations of the 2010 Eyjafjallajökull volcanic ash clouds. Atmos Environ 48:165–183. https://doi.org/10.1016/j.atmosenv.2011.06.072

Ganser GH (1993) A rational approach to drag prediction of spherical and nonspherical particles. Powder Technol 77(2):143–152. https://doi.org/10.1016/0032-5910(93)80051-B

Gilman MJ (1968) A brief survey of stopping rules in Monte Carlo simulations

Harvey NJ, Dacre HF, Webster HN et al (2020) The impact of ensemble meteorology on inverse modeling estimates of volcano emissions and ash dispersion forecasts: Grímsvötn 2011. Atmosphere 11(10):1022. https://doi.org/10.3390/atmos11101022

Hens AB, Tiwari MK (2012) Computational time reduction for credit scoring: an integrated approach based on support vector machine and stratified sampling method. Expert Syst Appl 39(8):6774–6781. https://doi.org/10.1016/j.eswa.2011.12.057

ICAO (2017) Roadmap for International Airways Volcano Watch (IAVW) in support of international air navigation - 11 December 2017, Version 3.0. Tech. rep., ICAO (International Civil Aviation Organisation) Meteorology Panel, Accessed on 19 Dec 2022

Jagers P (1986) Post-stratification against bias in sampling. Rev Int Sta 54(2):159–167. https://doi.org/10.2307/1403141

Jagers P, Odén A, Trulsson L (1985) Post-stratification and ratio estimation: usages of auxiliary information in survey sampling and opinion polls. Int Stat Rev 53(3):221–238. https://doi.org/10.2307/1402887

James PM (2006) An assessment of European synoptic variability in Hadley Centre Global Environmental models based on an objective classification of weather regimes. Clim Dyn 27(2–3):215–231. https://doi.org/10.1007/s00382-006-0133-9

James PM (2007) An objective classification method for Hess and Brezowsky Grosswetterlagen over Europe. Theor Appl Climatol 88(1–2):17–42. https://doi.org/10.1007/s00704-006-0239-3

Jenkins SF, Magill CR, McAneney K (2007) Multi-stage volcanic events: a statistical investigation. J Volcanol Geotherm Res 161(4):275–288. https://doi.org/10.1016/j.jvolgeores.2006.12.005

Jenkins SF, Magill C, McAneney J et al (2012) Regional ash fall hazard I: a probabilistic assessment methodology. Bull Volcanol 74(7):1699–1712. https://doi.org/10.1007/s00445-012-0627-8

Jenkins SF, Wilson TM, Magill C, et al. (2015) Volcanic ash fall hazard and risk. In: Global Volcanic Hazards and Risk. Cambridge University Press, p 173–222. 10.1017/CBO9781316276273.005

Jenkins SF, Biass S, Williams GT et al (2022) Evaluating and ranking Southeast Asia's exposure to explosive volcanic hazards. Nat Hazards Earth Syst Sci 22(4):1233–1265. https://doi.org/10.5194/nhess-22-1233-2022

Jones PD, Hulme M, Briffa KR (1993) A comparison of Lamb circulation types with an objective classification scheme. Int J Climatol 13(6):655–663. https://doi.org/10.1002/joc.3370130606

Kidson JW (2000) An analysis of New Zealand synoptic types and their use in defining weather regimes. Int J Climatol 20(3):299–316. https://doi.org/10.1002/(SICI)1097-0088(20000315)20:3<299::AID-JOC474>3.0.CO;2-B

Liu W, Bailey B (2002) Sample size determination for constructing a constant width confidence interval for a binomial success probability. Stat Probab Lett 56(1):1–5. https://doi.org/10.1016/S0167-7152(01)00029-3

Macedonio G, Costa A, Folch A (2008) Ash fallout scenarios at Vesuvius: numerical simulations and implications for hazard assessment. J Volcanol Geotherm Res 178(3):366–377. https://doi.org/10.1016/j.jvolgeores.2008.08.014

Marzocchi W, Selva J, Costa A, et al. (2015) Tephra fall hazard for the Neapolitan area. In: Loughlin SC, Sparks RSJ, Brown SK, Jenkins SF, Vye-Brown C (eds) Global volcanic hazards and risk. Cambridge University Press Cambridge, chap 6, p 239–248,https://dx.doi.org/10.1017/CBO9781316276273.008

Mastin LG, Guffanti M, Servranckx R et al (2009) A multidisciplinary effort to assign realistic source parameters to models of volcanic ash-cloud transport and dispersion during eruptions. J Volcanol Geotherm Res 186(1–2):10–21. https://doi.org/10.1016/j.jvolgeores.2009.01.008

Newhall CG, Self S (1982) The volcanic explosivity index (VEI) an estimate of explosive magnitude for historical volcanism. J Geophys Res Oceans 87(C2):1231–1238. https://doi.org/10.1029/JC087iC02p01231

ONR (2020)Underpinning the UK nuclear design basis criterion for naturally occurring external hazards final report. Tech. rep., Office for Nuclear Regulation, https://www.onr.org.uk/documents/2020/onr-rrr-059.pdf, Accessed on 20 Dec 2022

Robertson AW, Ghil M (1999) Large-scale weather regimes and local climate over the western United States. J Clim 12(6):1796–1813. https://doi.org/10.1175/1520-0442(1999)012%3C1796:LSWRAL%3E2.0.CO;2

Rougier JC (2013) Quantifying hazard losses. In: Rougier JC, Sparks RSJ, Hill LJ (eds). Risk and uncertainty assessment for natural hazards. Cambridge University Press, Chap 2, p 19–39. https://doi.org/10.1017/CBO9781139047562

Solman SA, Menéndez CG (2003) Weather regimes in the South American sector and neighbouring oceans during winter. Clim Dyn 21(1):91–104. https://doi.org/10.1007/s00382-003-0320-x

Suzuki T (1983) A theoretical model for dispersion of tephra. Arc Volcanism Phys Tectonics 95:113

Swindles GT, Lawson IT, Savov IP et al (2011) A 7000 yr perspective on volcanic ash clouds affecting northern Europe. Geology 39(9):887–890. https://doi.org/10.1130/G32146.1

Titos M, Martínez Montesinos B, Barsotti S et al (2022) Long-term hazard assessment of explosive eruptions at Jan Mayen (Norway) and implications for air traffic in the North Atlantic. Nat Hazards Earth Syst Sci 22(1):139–163. https://doi.org/10.5194/nhess-2021-264

Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22(158):209–212. https://doi.org/10.1080/01621459.1927.10502953

WMO-IUGG (2019) Conjoint session: seventh WMO VAAC "best practice" workshop (VAAC BP/7) and ninth WMO/IUGG volcanic ash scientific advisory group meeting (VASAG/9), Washington DC, United States of America, 21-22 November 2019. Tech. rep., World Meteorological Organization, International Union of Geodesy and Geophysics, https://old.wmo.int/aemp/sites/default/files/conjoint-vaac-bp-7-vasag-9_final-report.pdf, Accessed on 19 Dec 2022

Zidikheri MJ, Lucas C (2021) Improving ensemble volcanic ash forecasts by direct insertion of satellite data and ensemble filtering. Atmosphere 12(9):1215. https://doi.org/10.3390/ATMOS12091215

Authors and Affiliations

Jeremy Phillips[1] · Shannon Williams[2] · Anthony Lee[2] · Susanna Jenkins[3]

Shannon Williams
Shannon.Williams@bristol.ac.uk

Anthony Lee
Anthony.Lee@bristol.ac.uk

Susanna Jenkins
Susanna.Jenkins@ntu.edu.sg

1 School of Earth Sciences, University of Bristol, Bristol, UK

2 School of Mathematics, University of Bristol, Bristol, UK

3 Earth Observatory of Singapore, Asian School of the Environment, Nanyang Technological University, Singapore, Singapore