Two Ensemble Approaches for Forecasting Sulfur Dioxide Concentrations from Kīlauea Volcano

LACEY HOLLAND AND STEVEN BUSINGER

Department of Atmospheric Sciences, University of Hawai'i at Mānoa, Honolulu, Hawaii

TAMAR ELIAS

U.S. Geological Survey, Hawaiian Volcano Observatory, Hilo, Hawaii

TIZIANA CHERUBINI

Department of Atmospheric Sciences, University of Hawai'i at Mānoa, Honolulu, Hawaii

(Manuscript received 18 September 2019, in final form 13 June 2020)

ABSTRACT

Kīlauea volcano, located on the island of Hawaii, is one of the most active volcanoes in the world. It was in a state of nearly continuous eruption from 1983 to 2018 with copious emissions of sulfur dioxide (SO_2) that affected public health, agriculture, and infrastructure over large portions of the island. Since 2010, the University of Hawai'i at Mānoa provides publicly available vog forecasts that began in 2010 to aid in the mitigation of volcanic smog (or "vog") as a hazard. In September 2017, the forecast system began to produce operational ensemble forecasts. The months that preceded Kīlauea's historic lower east rift zone eruption of 2018 provide an opportunity to evaluate the newly implemented air quality ensemble prediction system and compare it another approach to the generation of ensemble members. One of the two approaches generates perturbations in the wind field while the other perturbs the sulfur dioxide (SO_2) emission rate from the volcano. This comparison has implications for the limits of forecast predictability under the particularly dynamic conditions at Kīlauea volcano. We show that for ensemble forecasts of SO_2 generated under these conditions, the uncertainty associated with the SO_2 emission rate approaches that of the uncertainty in the wind field. However, the inclusion of a fluctuating SO_2 emission rate has the potential to improve the prediction of the changes in air quality downwind of the volcano with suitable postprocessing.

1. Introduction

The longest eruptive episode in recorded history for Kīlauea volcano on the island of Hawaii ended in early August 2018. Kīlauea's recent episode was an effusive (nonexplosive) eruption that resulted in a continuous source of volcanic gas emissions for the 35 years since 1983. The eruption ended with eruptive fissures in Kīlauea's lower east rift zone (LERZ). The LERZ eruption was an extreme event that had impacts on visibility as far away as the Mariana Islands, more than 6000 km away (Guam Homeland Security 2018).

The prevailing northeasterly trade wind regime that dominates the weather of Hawaii advects volcanic emissions from Kīlauea to the southwest of the island chain. Volcanic emissions from Kīlauea can reach the other Hawaiiian islands when the predominant northeasterly trade wind regime is interrupted. Oahu, the most heavily populated island, regularly experienced aerosol impacts from Kīlauea, despite its location more than 300 km away. Volcanic emissions reach Oahu during episodes of southeasterly surface winds associated with precold frontal conditions, upper-level disturbances, and Kona low conditions (Tofte et al. 2017).

Although the most recent eruption has ended for Kīlauea, the state of Hawaii contains multiple active volcanoes that will erupt again in the near future, if past activity is any guide. In July 2019, the U.S. Geological Survey's (USGS) Hawaiian Volcano Observatory increased the alert level for Mauna Loa volcano to "advisory" (USGS 2019a). Historically, eruptions from Mauna Loa differ from those of Kīlauea in duration and vigor. During its 1984 eruption, emissions from

Corresponding author: Lacey Holland, lh33@hawaii.edu

DOI: 10.1175/WAF-D-19-0189.1

Mauna Loa affected most of the state (USGS 2019b). Future Mauna Loa eruptions are expected to repeat this pattern (Pattantyus et al. 2018a).

The noxious haze that arises from volcanic gas and aerosol emissions is known as "vog" or "volcanic smog." Vog originates from the gases dissolved within magma. As magma rises within the earth and approaches the surface, substantial amounts of gas are released into the atmospheric environment surrounding Kīlauea (Edmonds et al. 2013). Water vapor, carbon dioxide, sulfur dioxide (SO_2), hydrogen sulfide, and hydrogen halides are among the components of the exsolved gases (Mather et al. 2012).

Vog has appreciable and detrimental effects on human health (Longo et al. 2010; Tam et al. 2016), water quality, agriculture, infrastructure (Elias et al. 2009), and the local economy (Halliday et al. 2019). Vog is particularly harmful to those with respiratory conditions such as asthma, sinusitis, and respiratory diseases (Kleinman 1995; Ruben et al. 1995; Mannino et al. 1995; Worth 1995; Tam et al. 2007; Longo et al. 2010). Two EPA criteria air pollutants are among its components: SO_2 gas and fine particulate matter of size 2.5 μm or less ($PM_{2.5}$). The $PM_{2.5}$ component of vog is primarily sulfate aerosol (henceforth, SO_4) and only on rare occasion contains ash because of Kīlauea's effusive eruptive style. The SO_4 primarily forms as a secondary pollutant from the oxidation of SO_2.

Kīlauea was a significant source of SO_2 pollution. Between 2014 and 2017, Kīlauea averaged an SO_2 emission rate of nearly 2 million tons per year (Elias et al. 2018). This number far exceeds the 1.26 million tons of SO_2 emissions from all electricity generated in the United States during 2018 (Environmental Protection Agency 2019). As a potent, volcanic source of SO_2 within a remote, tropical environment, emissions from Kīlauea evolve differently than urban and industrial sources of SO_2. Pattantyus et al. (2018b) describe some of the complexities in the SO_4 pathways and estimate the rates of these reactions for Kīlauea.

Air quality forecasts and warnings can mitigate the public health risks associated with vog exposure. Efforts to develop vog forecasts at the University of Hawai'i at Mānoa (UHM) began in the 1990s to meet the need for air quality guidance within the state of Hawaii (Businger et al. 2015; Hollingshead et al. 2003). The transport and dispersion (hereafter "vog model") that predicts SO_2 and SO_4 concentrations within the state of Hawaii was developed initially as a proof-of-concept exercise.

The vog model is a forecast system that has operated nearly continuously since its implementation at UHM in 2011. Among its users is the Honolulu National Weather Service Forecast Office, which historically has used vog model guidance to forecast the reduced visibility that impacts the aviation and marine weather communities (R. Ballard 2018, personal communication). The vog model uses state-of-the-science regional weather forecasts with real-time volcanic SO_2 emission rates that are input to a custom implementation of a Lagrangian dispersion model [Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT; Draxler and Hess 1997; Stein et al. 2015)] to predict the impact of vog on local air quality for the state of Hawaii. It now includes a 27-member forecast ensemble that runs on a High-Performance Computing (HPC) cluster at UHM.

Attempts to develop air quality forecasts for vog without numerical weather prediction (NWP) guidance have achieved varying degrees of success. One study (Michaud et al. 2007) developed statistical relationships between meteorological variables to describe conditions that lead to poor air quality in Hawaii. The study suggested the impact of local wind patterns on the spatiotemporal variability of SO_2 near Kīlauea were too important to ignore. Others have since succeeded in applying statistical models that generally perform well at short lead times (less than 6 h) and show potential at longer lead times during steady, trade wind conditions (Reikard 2012).

These and other statistical forecast models that do not simulate physical processes directly, can benefit from the use of physical models (i.e., weather or air quality models) to improve forecast skill. Forecast techniques based on the leveraging statistical relationships are known to add the most value to short-term forecasts, while physical models generally perform better at longer lead times. Although comparisons have been made between these two types of models, efforts to develop physical or statistical models should be viewed as complementary because they can be applied together. Examples include model output statistics (MOS; Glahn and Lowry 1972), various ensemble MOS (Wilks and Hamill 2007) and analog ensemble methods (Delle Monache et al. 2013; Eckel and Delle Monache 2016), neural networks (Gardner and Dorling 1999), and blends thereof (Larson and Westrick 2006; Giorgi et al. 2011). This list is by no means comprehensive but merely demonstrates the broad range of statistical forecast techniques that can enhance the skill of physically based models or vice versa.

Among the findings of statistical forecast studies of vog, Reikard (2019) noted the challenges that nonstationarity in the mean and variance of SO_2 concentrations introduces. He also noted the difficulty in forecasting extreme vog events, which are the most critical events. Although nonstationarity is challenging to address, both problems can be approached with

stochastic (probabilistic) methods. Ensemble forecasts generally provide superior guidance during extreme events because they aim to produce a probability distribution that encompasses a range of likely outcomes.

Probability distributions, such as those that ensembles produce, are more useful to decision-makers than single deterministic forecast realizations (NRC 2006; AMS 2008; Gill et al. 2008; Hirschberg et al. 2011; Pattantyus and Businger 2015). Deterministic forecasts cannot adequately characterize the range of scenarios for which emergency planners and others need to prepare and do not directly provide the uncertainty information that WMO Guidelines recommend (WMO 2012). For these reasons, the vog model now produces operational ensemble forecasts.

The goal of an ensemble forecast is to provide a range and probability of scenarios that may occur because of the limitations inherent in an imperfect forecast system. The differences between scenarios are referred to as ''uncertainty.'' Ensemble forecasts approximate a probability distribution using a finite number of scenarios (Leith 1974). In a well-calibrated and unbiased ensemble forecast system, the expected outcome from this compilation of forecasts often more closely resembles the observed outcome than a single, deterministic forecast. This closer resemblance to observed outcomes explains the practice of using the ensemble mean itself as a forecast. When the ensemble forecast distribution is normal, its mean is the expected value of the forecast.

Forecast uncertainty commonly arises from either error in the characterization of initial conditions or from deficiencies in the model itself. More specific contributions to forecast error can be attributed to processes that fall within these broad categories: the contributions of model error, observation error, data assimilation procedures, and boundary conditions (Buizza et al. 2005). Most error contributions to air quality forecasts also fall under those same broad categories, with perhaps the addition of errors related to reactive chemistry mechanisms and rates (Delle Monache and Stull 2003), such as the partitioning between SO_2 and SO_4 in the vog model (Pattantyus et al. 2018b). These sources of uncertainty limit the predictive accuracy of models at longer forecast times through contributions to cumulative forecast error.

The wind-varying operational vog model ensemble simulates only one source of uncertainty in the initial conditions. It simulates the uncertainty from small errors in the initial wind field that contribute to the accumulation of errors in the HYSPLIT trajectories (Draxler 2003). As described by Pattantyus and Businger (2015) in their initial demonstration and qualitative assessment, each ensemble member simulates transport errors in the initial field through offsets (± 1 grid point) in the three-dimensions (x, y, z) of the modeled wind field. Although the wind-varying vog model attempts to characterize one source of uncertainty in the initial conditions (uncertainty ascribed to errors in the initial wind field) other sources of uncertainty also exist and may warrant inclusion in the vog model ensemble. There have been successes in other ensemble approaches that use multiphysics (Jiménez-Guerrero et al. 2013), multimodel (Delle Monache and Stull 2003), and postprocessing (Djalalova et al. 2015; Garner and Thompson 2013) approaches to ensemble modeling for air quality applications.

The second approach is novel in that it simulates the uncertainty associated with the active subsurface geology. The dynamics of the magma beneath Kīlauea govern the variations in the emission rate at Halema'uma'u (Kīlauea's summit) (Patrick et al. 2018) and include processes such as convection, degassing, and mixing that occur on rapid time scales (Edmonds et al. 2013). Preceding the summer 2018 eruption in the lower east rift zone, a lava lake at Kīlauea summit was visible at the surface and particularly active. FLYSPEC instruments positioned downstream of the summit provide an estimate of the variation in the SO_2 emission rate (Businger et al. 2015; Elias and Sutton 2017). From these data, we can estimate how much the varying emission rate contributes to vog forecast errors. These may also be compared to the magnitude of uncertainties that arise from transport errors in the initial wind field.

To quantify the amount of SO_2 ensemble forecast error that a varying emission rate contributes, we validate the current operational ensemble prediction system (hereafter, wind-varying ensemble) and compare it to the skill of an ensemble created by varying the emission rate. Although other sources of uncertainty exist and may affect the vog model ensemble, we focus on only the contributions from the varying emission rate at Kīlauea. We examine the performance of both ensembles using observations collected at the Pahala Hawaii Department of Health (HDOH) air quality monitoring station near the Kīlauea summit (Fig. 1). For this comparison, we examine SO_2 concentration forecasts during northeast trade wind conditions when the Pahala site is downwind of the Kīlauea summit.

2. Data and methods

To compare the skill of the wind-varying vog ensemble prediction system to the emission-varying ensemble, we analyzed the performance during the period from January to April 2018. Persistent trade wind conditions

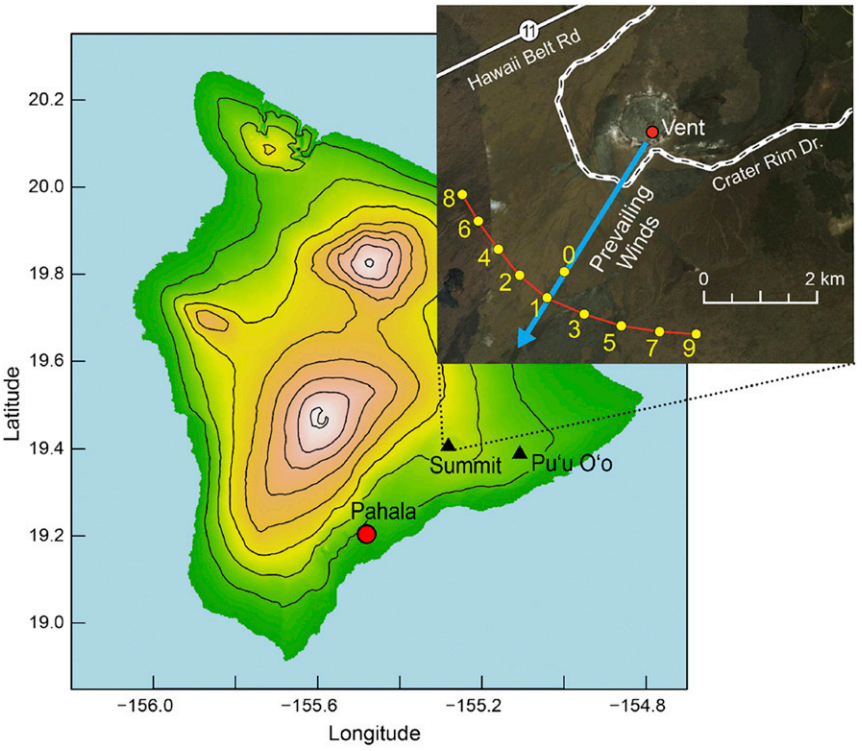

FIG. 1. Elevation contour map showing the locations of the Kīlauea summit and Pu'u 'Ō'ō vent (black triangles) and Pahala air quality monitor (red dot). The inset shows the location of the FLYSPEC instrument array southwest of the summit vent. The blue arrow shows the direction prevailing northeast trade winds.

were present over the Kīlauea summit for the 33 days included in the analysis. Under trade wind conditions, reliable estimates of the emission rate are available from the FLYSPEC array, located southwest of the summit vent. Moreover, trade winds advect vog toward the nearby HDOH air quality monitor in Pahala used for model validation (Fig. 1). Pahala is in an area ~30 km southwest of the summit emission source. Although SO_2 from Pu'u 'Ō'ō can directly impact the air quality measurements at Pahala, the emission rate and emission rate variability were one and two orders of magnitude smaller, respectively, than those from Halema'uma'u during the period of study.

In this study, we focus on forecasts of near-ground hourly averaged SO_2 concentrations at the Pahala HDOH monitoring site, without consideration of SO_4. This is because currently there is no direct measurement of SO_4 aerosol on Hawaii Island. The abundance of non-SO_4 sources of $PM_{2.5}$ (e.g., sea salt) impacts Pahala and makes it a challenge to validate the SO_4 forecasts directly (Businger et al. 2015). There is a large amount of variability in the SO_2 to SO_4 conversion rate (Porter and Clarke 1997; Kroll et al. 2015; Pattantyus et al. 2018b). For these reasons, we focus on forecasts of SO_2 concentrations at the nearby HDOH Pahala air quality monitor.

a. *Vog ensemble prediction system*

The "vog model" (Businger et al. 2015) is a custom version of the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT; Draxler and Hess 1997) model that is run operationally at UHM. The UHM implementation of HYSPLIT forecasts SO_2 gas and SO_4 aerosol concentrations that uses a fixed SO_2 to SO_4 conversion rate with active dry and wet deposition. The forecasts represent an hourly average of concentrations between 0 and 100 m above ground level. Current postprocessing applies appropriate air quality thresholds to indicate human health risk.

Meteorological input for the vog model comes from a custom Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW) with data assimilation that produces gridded forecasts twice per day. In addition to the use of the WRF-ARW dynamical core, the operational WRF implementation used for the vog model is unique and differs markedly from the NAM Hawaii nest. The custom WRF-ARW configuration contains 2 two-way nested domains, with

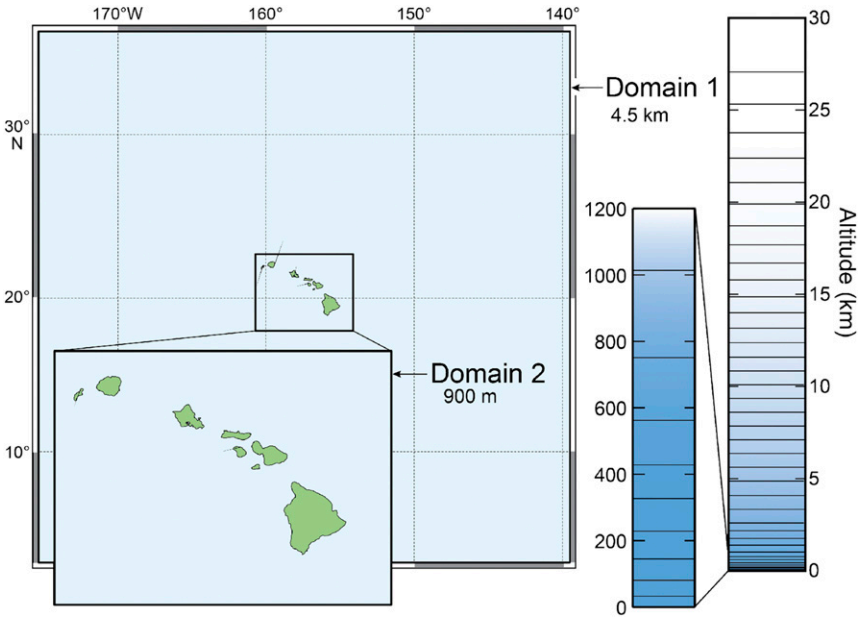

FIG. 2. Domains from the WRF-ARW that provide meteorological input to the vog model. The outer domain (domain 1) has 4.5-km resolution. The inner domain (domain 2) has 900-m resolution and includes 51 vertical levels with a fixed top near 40 hPa.

horizontal resolution spacing of 4.5 km and 900 m, extending over the central Pacific area and the island chain, respectively. One nested-down (from the coarser 4.5 km) domain of 900 m covers the Hawaiian island chain (Fig. 2). A total of 51 vertical levels are used. The vertical spacing is on the order of tens of meters for the levels nearest the ground, with first model level ~16 m above the surface, and gradually increases with height as shown in Fig. 2. The spacing never exceeds ~600 m between contiguous levels. The model top is fixed at 40 hPa, which corresponds to a height of ~22 km above ground level.

The WRF physics package that the current operational configuration uses includes (i) WRF single-moment 6-class scheme, which resolves ice, snow, and graupel processes suitable for high-resolution simulations (Hong and Lim 2006); (ii) the Mellor–Yamada–Janjić (MYJ) planetary boundary layer scheme (Janjić 2002), which solves the prognostic equation for the turbulent kinetic energy; (iii) the Rapid Radiation Transfer Model (RRTM) longwave–shortwave radiation scheme (Mlawer et al. 1997); and (iv) a simple downward integration that efficiently allows for clouds and clear-sky absorption and scattering (Dudhia 1989).

The WRF-ARW Model routinely runs four times daily with initial conditions at synoptic times (0000, 0600, 1200, and 1800 UTC) produced by a WRF data assimilation system. The WRF data assimilation system ingests local surface and upper-air observations along with aircraft and satellite observations. Boundary conditions are updated every six hours with model output from the National Centers for Environmental Prediction (NCEP) Global Forecasting System (GFS). Each forecast cycle produces a 60-h duration forecast that is output twice daily with forecast output in 3-hourly increments.

The operational vog model incorporates emission estimates of SO_2 from the USGS, as described in Businger et al. (2015). The initial emission rate is distributed among 20 sources, 10 each for the summit emission source and the Pu‘u ‘Ō‘ō vent. The sources are distributed as vertical line emissions with a tilt related to the prevailing trade winds. The largest portion of emissions is aloft. These parameters were based on prior empirical studies.

During the period of this study (January–April 2018), the two primary sources of volcanic SO_2 were the Kīlauea summit vent and the Pu‘u ‘Ō‘ō vent (Fig. 1). Emissions from the summit vent were measured at high temporal resolution using an array of FLYSPEC instruments located southwest of the summit and vent. The real-time emission rates that USGS provides represent a rate that has been averaged over nearly a week. The FLYSPEC array provides emission estimates at high temporal resolution (Horton et al. 2003, 2006, 2012; Elias et al. 2006; Elias and Sutton 2012; Elias et al. 2018; Businger et al. 2015) (Fig. 1 inset). The vog model ingests USGS weekly averaged emission rate estimates

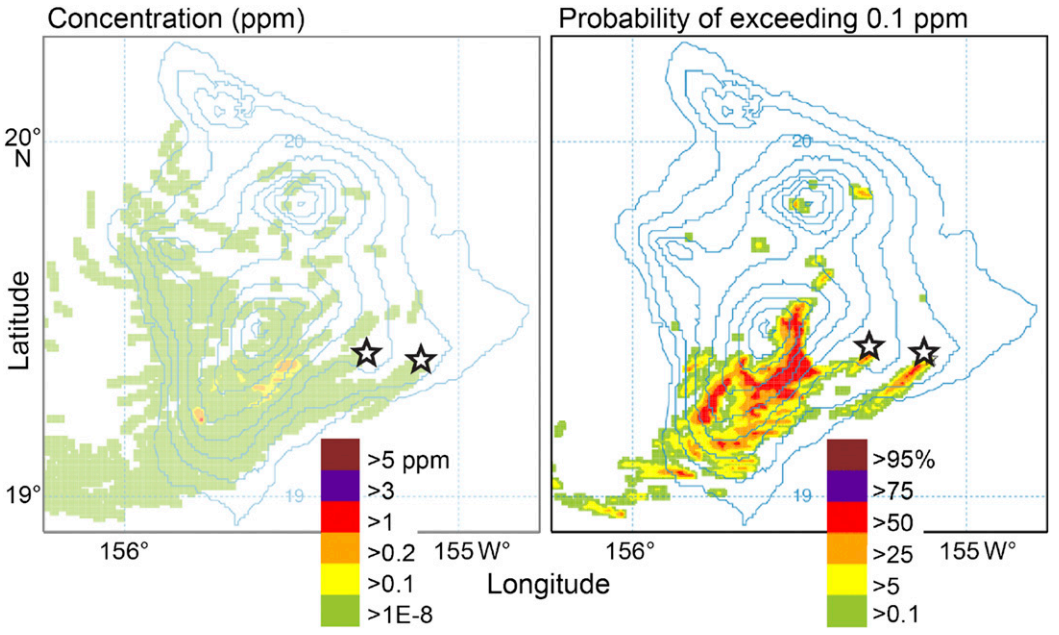

FIG. 3. Examples of (a) the operational deterministic forecast of SO_2 concentrations and (b) the probability that the SO_2 concentration will exceed 0.1 ppm are shown based on the wind-varying ensemble. The stars indicate the locations of the (left) Kīlauea summit and (right) Pu'u 'Ō'ō emission sources. Solid blue contour lines indicate elevation above sea level in 500-m increments.

from the FLYSPEC array to produce hourly SO_2 and SO_4 forecasts 60 h out twice per day (0000 and 1200 UTC). These forecasts are at 900-m horizontal resolution over the main Hawaiian Islands. An example of the deterministic forecast under prevailing trade wind weather conditions is in Fig. 3a.

In September 2017, an operational 27-member ensemble was implemented operationally. The ensemble uses the methods described in Pattantyus and Businger (2015) and Draxler (2003). This method perturbs the underlying wind field with variations in the meridional, zonal, and vertical directions for a total of 27 (3^3) combinations. Among other products such as the ensemble mean, the vog model ensemble produces forecasts of the probability that different air quality categories are exceeded (Fig. 3b). The concentrations and probabilities that are shown represent an average over a layer 0–100 m above ground level (AGL) based on the locations and characteristics of HYSPLIT-generated Lagrangian particles.

To examine how varying emission rates impact forecast uncertainty, an "emission-varying" ensemble was created. This ensemble differs from the operational (wind-varying) ensemble forecast in how its perturbations are generated. The wind-varying model perturbs the wind field and incorporates weekly averaged SO_2 emission rates.

To create the emission-varying ensemble, we resample hourly FLYSPEC emission rates from the previous day to simulate the variability in observed emission rates within the ensemble. The 10-s FLYSPEC emission rates, with a higher degree of quality control than available in the real-time weekly averages, were available for January–April 2018 and converted to hourly averages. The difference in quality control impacts the concentration forecasts and necessitates the use of appropriate quality assessment metrics when comparing the two ensembles.

Emission rates at the summit for each day demonstrate variability within the hourly averages and a skewed distribution (Fig. 4), so resampling methods are used. HYSPLIT makes use of fictitious "particles" (i.e., air parcels) to simulate atmospheric transport and dispersion. The locations of the particles for the emission-varying ensemble are initialized from the previous 12-h cycle using the same initial locations as the operational deterministic forecast. The next forecast cycle is run with a daily averaged emission rate and is used to initialize the emission-varying ensemble.

In addition to the 26 ensemble members based on resampled emission rates, the daily averaged emission rate is also used as an ensemble member for a total of 27 ensemble members, although the previous day's emission rates are oversampled, the result is the same number of members as the wind-varying ensemble that represents the distribution of emission rates.

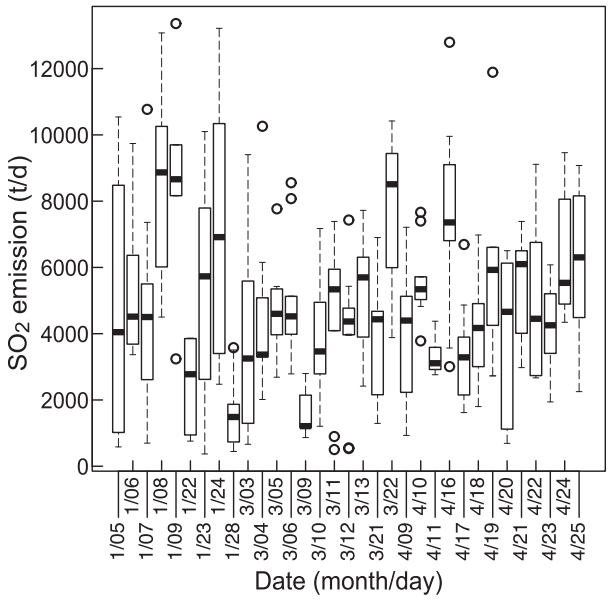

FIG. 4. Box-and-whisker plots show the distribution of hourly averaged SO_2 emission rates for days of persistent trade winds between January and April 2018. The center line indicates the median (Q2). The interquartile range (IQR), or difference between the first and third quartiles (Q1, Q3), is shown as the length of the boxes. Whiskers extend to the greater (lesser) of Q1 − 1.5 × IQR (Q3 + 1.5 × IQR) or the lowest (highest) emission rate contained therein. Outliers, values outside of the bounds of the whiskers, are shown as open circles.

b. Measurements

We use air quality measurements of SO_2 from the Hawaii Department of Health (HDOH) long-term monitors to assess the quality of vog model forecasts of SO_2 concentration. For this study, we use only days with prevailing northeasterly trade winds, or about 27.5% of days. These days were determined from anemometer-based wind measurements taken less than 2 km southwest of the summit emission source when the vector-averaged wind speed exceeds 4 m s^{-1}, and the wind direction was between 25° and 75°. The wind speed ranges were selected based on the annual wind speed distribution and the need to avoid measurement errors in wind direction at low wind speeds. The range of wind directions was based on the statistical major mode of the distribution of wind directions at the site. We also examine the wind characteristics at Pahala. The Pahala site has a pronounced diurnal sea breeze and mountain–valley circulation forced by differential radiation across the island of Hawaii and nearby ocean (Smith and Grubisic 1993) (Fig. 5) that necessitates the use of a mesoscale model that accurately characterizes these circulations.

c. Forecast performance analytics

To assess the skill of the vog model ensemble system, we examine ensemble error characteristics and the ensemble skill–spread relationship across forecast lead times. First, we compare the characteristics of errors associated with the operational wind-varying ensemble system to that of the emission-varying ensemble. This comparison is performed through the use of the operational model's deterministic forecast, and that of the operational wind-varying ensemble mean as performance benchmarks. The deterministic forecast is one member of the operational ensemble and a stand-alone forecast product. Then, we examine the performance of the spread–skill relationship of each ensemble as compared to its own ensemble mean. A well-calibrated ensemble forecast accurately simulates uncertainty in the expected value of the forecast. In this section, we show the error metric formulations used to assess each ensemble (or when appropriate, the operational deterministic forecast). These statistics are calculated for each ensemble (or deterministic) forecast for the same sets of observations.

The deterministic and ensemble forecast performance are compared by using the mean absolute error (MAE) and the continuous ranked probability score (CRPS) (Unger 1985; Hersbach 2000; Gneiting and Raftery 2004). The CRPS is considered a proper forecast score, meaning that a forecast score cannot gain an advantage through a forecast that differs from the expected forecast value. We use the MAE and CRPS in conjunction with one another to compare single-member forecasts to the probability distribution of the wind-varying and emission-varying ensembles. The CRPS formulation is identical to the MAE in cases where the ensemble consists of exactly one member, such as the deterministic operational forecast or the mean of an ensemble used as a forecast.

For a single forecast realization x_i, such as the deterministic forecast, with corresponding observation y_i, the MAE is summed over all forecast and observation pairs as follows [Eq. (1)]:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i|. \tag{1}$$

The MAE expresses the average size of forecast errors.

Because we seek a comparison between the performance of a single, deterministic forecast and the wind-varying and emission-varying ensembles, we use the CRPS. The CRPS is the integral of Brier scores at all possible thresholds h for the predictand (a continuous random variable). Observed values are denoted y.

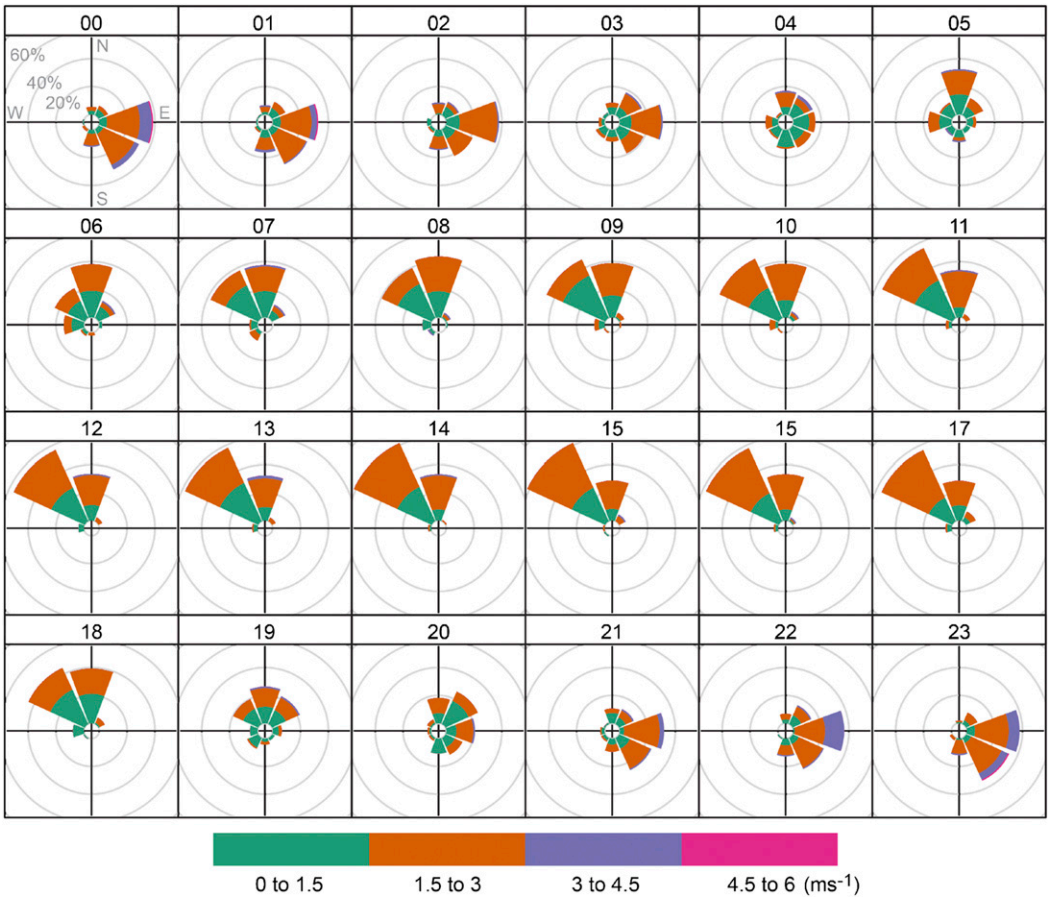

FIG. 5. Wind direction, speed (color), and frequency (%) for Pahala during January–April 2018 for each time of day (UTC; shown above each plot).

The CRPS is expressed in the following equations [Eqs. (2) and (3)]:

$$\mathrm{crps}(F, y) = \int_{-\infty}^{\infty} [F(h) - H(h - y)]^2 \, dt, \tag{2}$$

$$\mathrm{CRPS} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{crps}(F_i, y_i). \tag{3}$$

In this formulation, F is the predictive cumulative distribution function (CDF), or the distribution of ensemble forecast values. The term $H(h - y)$ is the Heaviside (or unit step) function. The Heaviside function has a value equal to zero when $h < y$ and has a value equal to one elsewhere. Because the CRPS is a generalized form of the MAE, this metric allows the direct comparison of skill between ensemble and deterministic forecasts.

The bias (mean error) [Eq. (4)] and the Spearman rank correlation are indicators of forecast quality. The bias shows the average (denoted with an overbar) differences between the corresponding observation x and forecast y pairs. It indicates if the forecast and observed concentrations are similar in size and if one generally reports a higher or lower value than the other, on average. The bias formulates as follows:

$$\mathrm{bias} = (\bar{y} - \bar{x}). \tag{4}$$

The bias is included to examine systemic differences between individual forecasts. In other words, this metric demonstrates if a forecast is usually higher (bias > 0) or lower (bias < 0) than the observations.

The Spearman rank correlation ρ shown in Eq. (5) indicates the correspondence between forecasts and observations. The Spearman rank is a robust, nonparametric indication of correlation. We focus on only comparing the physical processes that lead to vog forecast uncertainty and not necessarily in the strength of a linear relationship. Neither ensemble has undergone calibration or postprocessing which would improve the statistical relationship between the observations and ensemble forecasts:

$$\rho = \frac{6\sum d_i^2}{n(n^2-1)}, \tag{5}$$

where d_i indicates differences between the ranks of the forecast and observations and n is the number of observations. This statistic applies to the deterministic forecast and the mean of each ensemble to determine how well the most common way to use each ensemble corresponds to observations.

Mason and Weigel (2009) first developed the generalized discrimination score D. The generalized discrimination score quantifies the ability of a forecast to discern between different observed outcomes as one of the many dimensions of forecast quality (Murphy 1991). This generalized discrimination score is also known as the "two-alternative forced-choice" (2AFC) and quantifies the forecast attribute of "discrimination," the ability to discern between differences in observational values. Later, Weigel and Mason (2011) extend the formulation to include ensemble forecasts of continuous variables [Eq. (6)]. In this formulation, x denotes observations and y denotes ensemble forecasts. In this case, y_s and y_t denote two ensemble forecasts compared over n observations:

$$D = \frac{1}{2}(\tau_{\mathbf{R},\mathbf{x}} + 1). \tag{6}$$

In this formulation of D, $\tau_{\mathbf{R},\mathbf{x}}$ is Kendall's rank correlation coefficient (Sheshkin 2007) for n observations within the n-element vector of corresponding ensemble ranks $\mathbf{R} = (R_1, \ldots, R_n)$; $\mathbf{R}$ is the full n-element vector of ensemble ranks for all ensemble forecasts, which include y_s and y_t. The rank for ensemble forecast y_s within a set of n ensemble forecasts $(y_1, y_2, \ldots, y_n)$ is expressed in Eq. (7):

$$R_s = 1 + \sum_{\substack{t=1 \\ t \neq s}}^{n} u_{s,t}, \tag{7}$$

with $u_{s,t} = 1$, if $F_{s,t} > 0.5$; $u_{s,t} = 0.5$, if $F_{s,t} = 0.5$; and $u_{s,t} = 0$, if $F_{s,t} < 0.5$.

The term $F_{s,t}$ denotes the proportion of ensemble member pairs that exceed the threshold. The formulation for $F_{s,t}$ in this example is as formulated in Eq. (8):

$$F_{s,t} = \frac{\sum_{i=1}^{m_s} r_{s,t,i} - \dfrac{m_s(m_s+1)}{2}}{m_s m_t}, \tag{8}$$

where m_s is the number of ensemble members of y_s, and likewise for m_t. The rank of $y_{s,i}$ is denoted as $r_{s,t,i}$ with respect to the entire set of pooled ensemble members sorted in ascending order.

The derivation of D for continuous ensemble forecasts in Weigel and Mason (2011) is thorough, and these equations result in a measure that uses ranks to determine how well differences in ensemble forecasts distinguish between differences in observed values. A generalized discrimination score D of 0.5 denotes a forecast with no skill. Scores greater than 0.5 denote a skillful forecast. This score is also related to the area under the relative operating characteristic curve, which relates the hit rate and false alarm rate for multiple thresholds to indicate forecast skill (Buizza and Palmer 1998; Mason and Graham 1999).

Ensemble forecasts can characterize uncertainties in the SO_2 forecast that arise from model error, also called the spread–skill relationship. We evaluate the spread–skill relationship in terms of how well the ensemble spread characterizes errors in the ensemble mean forecast, as described in Hopson (2014). The mean absolute deviation of ensemble members from the ensemble mean (MAD_{EM}) is as follows [Eq. (9)]:

$$\mathrm{MAD}_{\mathrm{EM}} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|. \tag{9}$$

In this formulation, x_i is the value associated with an individual ensemble member, $\overline{x}$ is the value of the ensemble mean, and n is the number of ensemble members. The MAD_{EM} can be compared directly to the MAE for the ensemble mean to indicate how well the ensemble spread characterizes the errors in the ensemble mean. Thus, the MAE is calculated for the ensemble mean of the wind-varying and emission-varying ensembles and compared to the MAD_{EM} of each. In a well-calibrated ensemble, the larger errors in the MAE should correspond to a larger spread between ensemble members – and a larger MAD_{EM}.

3. Results

We compare the performance of the wind-varying ensemble to that of the operational deterministic forecast and the emission-varying ensemble. We first find a baseline for error magnitudes of SO_2 forecasts, and then examine measures of correspondence and the spread–skill relationship. We examine the error magnitudes for the operational deterministic forecast, the wind-varying (operational) ensemble, and the emission-varying ensemble. We show the skill of each ensemble mean. We then examine how well the deterministic forecast and ensemble means correspond to measurements of SO_2. Finally, we look at forecast discrimination, the ability of the forecasts to discern variations in SO_2 concentration.

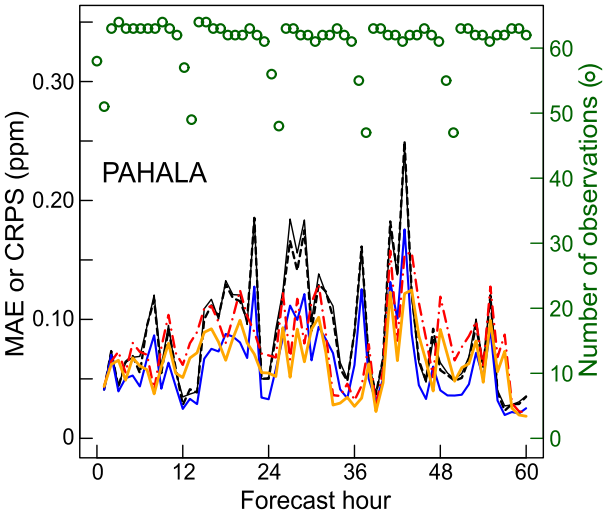

FIG. 6. The mean absolute error (MAE) for the deterministic forecast is shown (solid black line) for forecast lead hours 1–60. The MAE of the wind-varying ensemble mean is shown (black dashed line) with the continuous ranked probability score (CRPS) for the wind-varying ensemble (solid blue line). These are compared to the MAE for the experimental emission-varying ensemble mean (red dash–dotted line) and the CRPS for the experimental emission-varying ensemble (solid gold line). The number of observations for each forecast hour is shown (green dots) and corresponds to the axis on the right.

We examine the error magnitudes (MAE) of the operational deterministic forecast, the mean of the operational ensemble, and mean of the emission-varying ensemble to the CRPS of the wind-varying ensemble and the emission-varying ensemble.

Our results show the wind-varying ensemble mean has smaller errors than those of the deterministic forecast (Fig. 6). When the entire PDF of the wind-varying ensemble forecast is used, the CRPS indicates it exhibits smaller errors than using its mean as a forecast or the using the deterministic forecast upon which it is based.

The mean of the emission-varying ensemble has the most substantial errors of all forecasts in the comparison. The emission-varying ensemble as a whole displays errors roughly the same size as the deterministic operational forecast.

Forecast bias is the average difference between the forecast and observation (i.e., whether the forecasts are on average higher or lower than the observations). It indicates systematic errors. The mean of the emission-varying ensemble shows the most substantial bias at most lead times and forecasts concentrations that are much higher than observed (Fig. 7). The mean of the wind-varying ensemble displays the bias closest to zero for nearly all lead times and more closely matches the values of the observations than the deterministic forecast.

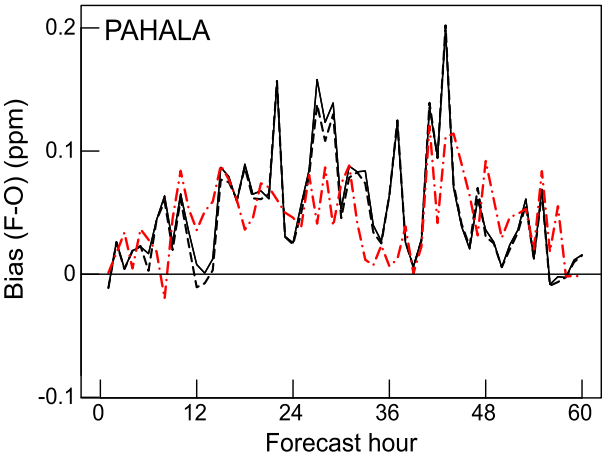

FIG. 7. Bias is shown for the operational deterministic forecast (solid black line), the wind-varying ensemble mean (dashed black line), and the mean of the emission-varying ensemble (dash–dotted red line). A thin, solid black line indicates the zero bias line.

We use the Spearman rank correlation to show how the operational deterministic, mean of the wind-varying ensemble, and the mean of the emission-varying ensemble correspond to measurements of SO_2 at Pahala. With this metric, we find the SO_2 concentrations simulated with the emission-varying ensemble generally correspond more closely to observed SO_2 concentrations than either the deterministic operational forecast or the wind-varying ensemble mean forecast (Fig. 8). This also has implications for the bias shown in Fig. 7. Because the correspondence is stronger, a greater portion of the error in the emission-varying ensemble are

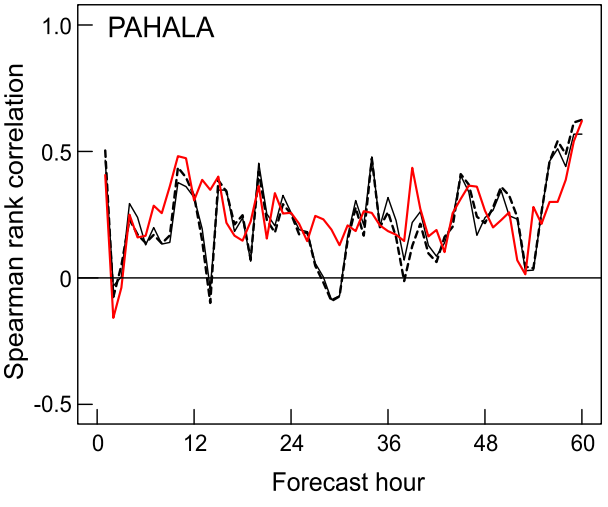

FIG. 8. The Spearman rank correlation is shown for the operational deterministic forecast (solid black line), the mean of the operational ensemble (dashed black line), and the mean of the emission-varying ensemble (solid red line) for forecast hours 1–60.

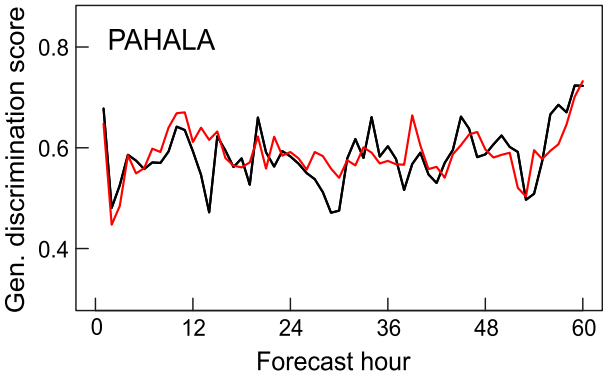

FIG. 9. The generalized discrimination score for ensembles is shown for forecast hours 1–60. The operational (wind-varying) ensemble (black line) is shown with the emission-varying ensemble (red line).

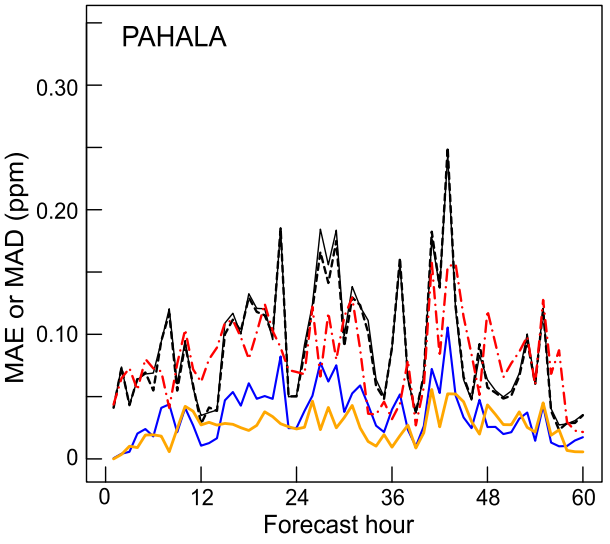

FIG. 10. The mean absolute error (MAE) for the deterministic forecast is shown (solid black line). The MAE of the ensemble mean forecast for the wind-varying ensemble is shown (black dashed line) with the mean absolute deviation (MAD) for the wind-varying ensemble spread (solid blue line). These are compared to the MAE for the emission-varying ensemble (red dash-dotted line) and the MAD for the spread of the emission-varying ensemble (solid gold line).

likely to be systematic relative to the portion of random errors than those observed in the other forecasts shown. Although the deterministic operational forecasts and mean of the wind-varying ensemble have smaller errors on average, portions of these errors arise from variability in the emission rate. Some of the error that emission rate variability introduces can be simulated.

The generalized discrimination scores show that the emission-varying ensemble has a better ability to distinguish between different observations of SO_2 (Fig. 9). The emission-varying ensemble is a skillful forecast for nearly all forecast hours and has a discrimination score that exceeds 0.5, the "no skill" value for this verification metric. During most forecast hours, discrimination score of the emission-varying ensemble is greater than that of the wind-varying ensemble. This is likely because during periods of eruptive activity, such as when there is an active lava lake is at Kīlauea's summit, the forecasts benefit from updated emission rates and variability information.

Another critical aspect of a forecast ensemble is the spread–skill relationship. With the MAE as the primary error metric, we compare the error magnitudes for the deterministic operational forecast, the wind-varying ensemble mean, and the mean of the emission-varying ensemble. For a well-calibrated ensemble, the spread should approximate the size of the errors of the ensemble forecast mean.

The spread of the wind-varying ensemble closely resembles the errors associated with the forecast mean and incurs smaller MAEs than the deterministic forecast (Fig. 10). The errors in the mean of the emission-varying ensemble are larger than its spread and larger than those of the deterministic forecast. There are indications that this relates to the large, systematic bias in the emission-varying ensemble (i.e., the emission-varying ensemble forecasts an excessive amount of vog) (Fig. 7). The emission-varying ensemble is unable to characterize the full range of errors in its forecast mean. There are additional sources of error for which the emission-varying ensemble does not account. The errors in the mean of the emission-varying ensemble are also larger than that of the wind-varying ensemble. However, the amount of spread in the emission-varying ensemble is about the same size as the spread in the wind-varying ensemble. This shows that the variability in the emission source can contribute as much to the variability in the forecast concentrations as variability in the wind near the emission source. There are also indications that the systematic bias present in the emission-varying ensemble contributes to the large errors in its ensemble mean.

Overall, the emission-varying ensemble that we present has larger errors than the wind-varying ensemble. These errors are similar in size to the operational deterministic forecast. The wind-varying ensemble and its mean have the lowest errors of all forecasts examined. The probability distribution of SO_2 forecasts in the wind-varying ensemble aptly characterizes the expected size of errors in its mean, when used as a forecast. However, the emission-varying ensemble had a better degree of correspondence to the measurements at Pahala and was

better able to distinguish between different observations at Pahala.

4. Discussion

We explored two different ways to create ensemble forecasts to predict SO_2 concentrations based on perturbing the initial conditions. One method is based on perturbations in the initial wind field (wind-varying ensemble) near the emission sources. The other is based on perturbations in the strength of the SO_2 emission source (emission-varying ensemble). The latter uses a relatively simple method to generate an ensemble that emulates forecast uncertainty from a variable SO_2 emission source.

Although the ensemble created from perturbations in the wind field led to a better forecast in terms of metrics related to error magnitude, the mean of the emission-varying ensemble had a better correlation with observations and an improved ability to discern between different observations of SO_2.

The emission-varying ensemble displayed a better discrimination score. This means that the emission-varying ensemble has a better ability to distinguish between differences in observations than the wind-varying ensemble. Both ensembles would benefit greatly from the use of postprocessing and other techniques that reduce systematic biases in the forecasts.

Bulk error statistics show the mean of the emission-varying ensemble has larger errors than both the mean of the wind-varying ensemble and the operational deterministic forecast. The weekly averages used in the wind-varying ensemble were received in real time without postprocessing review and thus received a different degree of quality control than the emission rates used to create the emission-varying ensemble. Prior efforts to bias-correct deterministic vog model forecasts were implemented using the weekly averaged emission rates. It is for this reason that the biases in the deterministic forecast and wind-varying ensemble mean are relatively small. The emission-varying ensemble did not have a similar bias-correction applied and such a treatment likely would affect its performance. It is for this reason that we emphasize the comparisons between the ensemble spread and measures of its discrimination.

There is potential to improve ensemble forecasts of SO_2 through the inclusion of an emission rate that varies. We find that the subweekly emission rate at Kīlauea varies to such a large extent that it contributes nontrivially to the fluctuations in SO_2 concentrations observed at Pahala. When we include the size of the fluctuations in the emission rate, the ability of the ensemble to predict changes in the air quality at Pahala improves.

Other studies have shown that ensemble skill may be improved through the characterization of additional processes or additional ensembles–if the members are skill ful, especially in cases where the ensemble is underdispersive (i.e., systematically fails to encompass the full range of likely outcomes) (Ebert 2001). Because the emission-varying ensemble is a skillful forecast, it could be used to increase the spread of the wind-varying vog ensemble. However, the current wind-varying ensemble does not appear to be underdispersive. Both methods to generate a vog ensemble are likely to benefit greatly from the application of postprocessing or statistical methods, such as the analog ensemble technique (Delle Monache et al. 2013).

Our study also suggests that for Kīlauea the variability in the emission rate itself places a nonnegligible constraint on the predictability of SO_2 concentrations. The uncertainty that arises from a varying emission rate is unique from, and of similar magnitude to, the uncertainty that arises from initial errors in atmospheric transport.

5. Conclusions

We compare the wind-varying vog model ensemble to an ensemble created by perturbing the strength of the volcanic SO_2 emission source. Our findings show that the emission-varying ensemble and its mean have substantially larger errors than the operational (wind-varying) forecasts. The reasons for this are not clearly understood and warrant further exploration. The emission-varying ensemble, however, is more skillful at discerning between relative concentrations of SO_2 as a proxy for vog (i.e., when more or less vog than the usual amount is expected.) This also means the emission-varying ensemble is likely to be a useful forecast for applications in which identifying relative amounts of SO_2 are helpful.

The emission-varying ensemble is able to simulate forecast uncertainty related to the magnitude of variations observed in the emission rate. The emission-varying ensemble is a skillful forecast that is likely to benefit from postprocessing and other methods that address systematic forecast biases. We also show that variability in the emission rate can produce nearly as much variation in the resulting SO_2 concentration forecast as perturbations in the initial wind field. Variability in the emission rate is likely to be a limiting factor in the predictability of concentrations of volcanic SO_2.

We show that when we include fluctuations in SO_2 emission rate, we are better able to predict the changes in air quality that occur downwind of the source. This relatively simple method to generate an ensemble shows that a varying emission rate can introduce nontrivial

amounts of forecast uncertainty in volcanic air pollution forecasts.

Acknowledgments. The work is funded through the Hawaii Department of Health (HDOH) Grant HEER-UH-ORS 2018, the Hawai'i Cane and Sugar Company Grant 127-4640-4, U.S. Geological Survey's Volcano Hazards Program Grant (127-4640-4), and the Office of Naval Research (ONR) Award 618 N00014-18-1-2166. We express our gratitude to the University of Hawai'i at Mānoa ITS HPC cluster computing services, and the many others who contributed to this work through discussion or support. We also express our gratitude to the two reviewers who provided insightful comments that aided this study.

REFERENCES

AMS, 2008: Enhancing weather information with probability forecasts: An information statement of the American Meteorological Society. *Bull. Amer. Meteor. Soc.*, **89**, 1049–1053.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518, https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Businger, S., R. Huff, K. Horton, A. J. Sutton, and T. Elias, 2015: Observing and forecasting vog dispersion from Kīlauea Volcano, Hawai'i. *Bull. Amer. Meteor. Soc.*, **96**, 1667–1686, https://doi.org/10.1175/BAMS-D-14-00150.1.

Delle Monache, L., and R. Stull, 2003: An ensemble air-quality forecast over western Europe during an ozone episode. *Atmos. Environ.*, **37**, 3469–3474, https://doi.org/10.1016/S1352-2310(03)00475-8.

——, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Djalalova, I., L. Monache, and J. Wilczak, 2015: PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmos. Environ.*, **108**, 76–87, https://doi.org/10.1016/j.atmosenv.2015.02.021.

Draxler, R. R., 2003: Evaluation of an ensemble dispersion calculation. *J. Appl. Meteor.*, **42**, 308–317, https://doi.org/10.1175/1520-0450(2003)042<0308:EOAEDC>2.0.CO;2.

——, and G. D. Hess, 1997: Description of the HYSPLIT_4 modeling system. NOAA Tech. Memo. ERL ARL-224, NOAA/Air Resources Laboratory, Silver Spring, MD, 27 pp., www.arl.noaa.gov/documents/reports/arl-224.pdf.

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Ebert, E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and L. Delle Monache, 2016: A hybrid NWP-analog ensemble. *Mon. Wea. Rev.*, **144**, 897–911, https://doi.org/10.1175/MWR-D-15-0096.1.

Edmonds, M., and Coauthors, 2013: Magma storage, transport and degassing during the 2008–10 summit eruption at Kīlauea Volcano, Hawai'i. *Geochim. Cosmochim. Acta*, **123**, 284–301, https://doi.org/10.1016/j.gca.2013.05.038.

Elias, T., and A. J. Sutton, 2012: Sulfur dioxide emission rates from Kīlauea Volcano, Hawai'i, 2007–2010. USGS Open File Rep. 2012-1107, 25 pp., http://pubs.usgs.gov/of/2012/1107/.

——, and ——, 2017: Volcanic air pollution hazards in Hawaii. U.S. Geological Survey Fact Sheet 2017–3017, 4 pp., https://doi.org/10.3133/fs20173017.

——, ——, C. Oppenheimer, K. A. Horton, H. Garbeil, V. Tsanev, A. J. S. McGonigle, and G. Williams-Jones, 2006: Comparison of COSPEC and two miniature ultraviolet spectrometer systems for SO_2 measurements using scattered sunlight. *Bull. Volcanol.*, **68**, 313–322, https://doi.org/10.1007/s00445-005-0026-5.

——, ——, J. P. Kauahikaua, J. D. Ray, and J. L. Babb, 2009: Ambient air quality effects of the 2008–2009 Halema'uma'u eruption on the Island of Hawai'i. *Eos, Trans. Amer. Geophys. Union*, **90** (Fall Meeting Suppl.), Abstract V43G-2337.

——, C. Kern, K. Horton, A. Sutton, and H. Garbeil, 2018: Measuring SO_2 emission rates at Kīlauea Volcano, Hawaii, using an array of upward-looking UV spectrometers, 2014–2017. *Front. Earth Sci.*, **6**, 214, https://doi.org/10.3389/feart.2018.00214.

Environmental Protection Agency, 2019: EPA releases 2018 power plant emissions demonstrating continued progress. Accessed 24 July 2019, https://www.epa.gov/newsreleases/epa-releases-2018-power-plant-emissions-demonstrating-continued-progress.

Gardner, M. W., and S. R. Dorling, 1999: Neural network modelling and prediction of hourly NO_X and NO_2 concentrations in urban air in London. *Atmos. Environ.*, **33**, 709–719, https://doi.org/10.1016/S1352-2310(98)00230-1.

Garner, G., and A. Thompson, 2013: Ensemble statistical post-processing of the national air quality forecast capability: Enhancing ozone forecasts in Baltimore, Maryland. *Atmos. Environ.*, **81**, 517–522, https://doi.org/10.1016/j.atmosenv.2013.09.020.

Gill, J., and Coauthors, 2008: Guidelines on communicating forecast uncertainty. WMO Tech. Doc. 4122, 25 pp.

Giorgi, M., A. Ficarella, and M. Tarantino, 2011: Assessment of the benefits of numerical weather predictions in wind power forecasting based on statistical methods. *Energy*, **36**, 3968–3978, https://doi.org/10.1016/j.energy.2011.05.006.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., and A. E. Raftery, 2004: Strictly proper scoring rules, prediction, and estimation. Tech. Rep. 463, Department of Statistics, University of Washington, 29 pp., www.stat.washington.edu/tech.reports.

Guam Homeland Security, 2018: Volcanic haze reaches the Marianas; those with respiratory issues advised to take precautions. Accessed 29 May 2019, https://www.ghs.guam.gov/volcanic-haze-reaches-marianas-those-respiratory-issues-advised-take-precautions.

Halliday, T. J., J. Lynham, and Á. de Paula, 2019: Vog: Using volcanic eruptions to estimate the health costs of particulates. *Econ. J.*, **129**, 1782–1816, https://doi.org/10.1111/ecoj.12609.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*,

15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hirschberg, P. A., and Coauthors, 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bull. Amer. Meteor. Soc.*, **92**, 1651–1666, https://doi.org/10.1175/BAMS-D-11-00073.1.

Hollingshead, A., S. Businger, R. Draxler, J. Porter, and D. Stevens, 2003: Dispersion modeling of the Kilauea plume. *Bound.-Layer Meteor.*, **108**, 121–144, https://doi.org/10.1023/A:1023086823088.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. *Mon. Wea. Rev.*, **142**, 1125–1142, https://doi.org/10.1175/MWR-D-12-00111.1.

Horton, K. A., J. Porter, P. Mouginis-Mark, C. Oppenheimer, and H. Garbeil, 2003: Apparatus for measuring radiation and method of use. U.S. Patent 7,148,488.

——, G. Williams-Jones, H. Garbeil, T. Elias, A. J. Sutton, P. Mouginis-Mark, J. N. Porter, and S. Clegg, 2006: Real-time measurement of volcanic SO_2 emissions: Validation of a new UV correlation spectrometer (FLYSPEC). *Bull. Volcanol.*, **68**, 323–327, https://doi.org/10.1007/s00445-005-0014-9.

——, H. Garbeil, A. J. Sutton, T. Elias, and S. Businger, 2012: Early monitoring results from the Halema'uma'u vog measurement and prediction FLYSPEC array. Extended Abstracts, *AGU Chapman Conf. on Hawaiian Volcanoes: From Source to Surface*, Waikoloa, HI, Amer. Geophys. Union, TH-34, http://hilo.hawaii.edu/~kenhon/HawaiiChapman/documents/1HawaiiChapmanAbstracts.pdf.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso Model. NCEP Office Note 437, 61 pp.

Jiménez-Guerrero, P., S. Jerez, J. P. Montávez, and R. M. Trigo, 2013: Uncertainties in future ozone and PM10 projections over Europe from a regional climate multiphysics ensemble. *Geophys. Res. Lett.*, **40**, 5764–5769, https://doi.org/10.1002/2013GL057403.

Kleinman, M. T., 1995: Health effects of inhaled particles and acid sulfate aerosols. *Proc. Vog and Laze Seminar*, Honolulu, HI, Hawaii State Department of Health.

Kroll, J. H., and Coauthors, 2015: Atmospheric evolution of sulfur emissions from Kilauea: Real-time measurements of oxidation, dilution, and neutralization within a volcano plume. *Environ. Sci. Technol.*, **49**, 4129–4137, https://doi.org/10.1021/es506119x.

Larson, K., and K. Westrick, 2006: Short-term wind forecasting using off-site observations. *Wind Energy*, **9**, 55–62, https://doi.org/10.1002/we.179.

Leith, C. E., 1974: Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Longo, B. M., W. Yang, J. B. Green, F. L. Crosby, and V. L. Crosby, 2010: Acute health effects associated with exposure to volcanic air pollution (vog) from increased activity at Kilauea Volcano in 2008. *J. Toxicol. Environ. Health*, **73A**, 1370–1381, https://doi.org/10.1080/15287394.2010.497440.

Mannino, D. M., S. M. Ruben, and F. C. Holschuh, 1995: Weekly variability of emergency room visits for asthma in Hilo, Hawai'i, 1981–1991. *Proc. Vog and Laze Seminar*, Honolulu, HI, Hawaii State Department of Health.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

——, and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349, https://doi.org/10.1175/2008MWR2553.1.

Mather, T. A., and Coauthors, 2012: Halogens and trace metal emissions from the ongoing 2008 summit eruption of Kilauea volcano, Hawaii. *Geochim. Cosmochim. Acta*, **83**, 292–323, https://doi.org/10.1016/j.gca.2011.11.029.

Michaud, J. D., J.-P. Michaud, and D. Krupitsky, 2007: Temporal variability in SO_2 exposures at Hawai'i Volcanoes National Park, USA. *Environ. Geol.*, **52**, 81–92, https://doi.org/10.1007/s00254-006-0459-y.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601, https://doi.org/10.1175/1520-0493(1991)119<1590:FVICAD>2.0.CO;2.

NRC, 2006: *Completing the Forecasts: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Academies Press, 124 pp.

Patrick, M. R., T. R. Orr, D. A. Swanson, T. Elias, and B. Shiro, 2018: Lava lake activity at the summit of Kīlauea Volcano in 2016. U.S. Geological Survey Scientific Investigations Rep. 2018–5008, 58 pp., https://doi.org/10.3133/sir20185008.

Pattantyus, A., and S. Businger, 2015: Ensemble forecasting of volcanic emissions in Hawai'i. *Ann. Geophys.*, **57**, https://doi.org/10.4401/ag-6607.

——, L. Holland, S. Businger, and T. Elias, 2018a: Projecting air quality impacts for the next eruption of Mauna Loa Volcano, Hawai'i. *20th Conf. on Air Chemistry*, Austin, TX, Amer. Meteor. Soc., 7.3A, https://ams.confex.com/ams/98Annual/webprogram/Paper335864.html.

——, S. Businger, and S. Howell, 2018b: Review of sulfur dioxide to sulfate aerosol chemistry at Kīlauea Volcano, Hawai'i. *Atmos. Environ.*, **185**, 262–271, https://doi.org/10.1016/j.atmosenv.2018.04.055.

Porter, J., and A. Clarke, 1997: Aerosol size distribution models based on in situ measurements. *J. Geophys. Res.*, **102**, 6035–6045, https://doi.org/10.1029/96JD03403.

Reikard, G., 2012: Forecasting volcanic air pollution in Hawai'i: Tests of time series models. *Atmos. Environ.*, **60**, 593–600, https://doi.org/10.1016/j.atmosenv.2012.06.040.

——, 2019: Volcanic emissions and air pollution: Forecasts from time series models. *Atmos. Environ.: X*, **1**, 100001, https://doi.org/10.1016/j.aeaoa.2018.100001.

Ruben, S. M., D. M. Mannini, F. C. Holschuh, T. C. Holshuh, M. D. Wilson, and T. Holschuh, 1995: Emergency room visits for asthma and chronic obstructive pulmonary disease on the Island of Hawai'i, 1981–1991. *Proc. Earthquake, Tsunami, and Volcano Hazards Seminar*, Hilo, HI, University of Hawaii at Hilo.

Sheshkin, D. J., 2007: *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 1776 pp.

Smith, R. B., and V. Grubisic, 1993: Aerial observations of Hawaii's wake. *J. Atmos. Sci.*, **50**, 3728–3750, https://doi.org/10.1175/1520-0469(1993)050<3728:AOOHW>2.0.CO;2.

Stein, A. F., R. R. Draxler, G. D. Rolph, B. J. B. Stunder, M. D. Cohen, and F. Ngan, 2015: NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bull. Amer. Meteor. Soc.*, **96**, 2059–2077, https://doi.org/10.1175/BAMS-D-14-00110.1.

Tam, E., and Coauthors, 2007: Volcanic air pollution and respiratory symptoms in school children on the Big Island of Hawai'i. *Proc. ATS 2007*, San Francisco, CA, American Thoracic Society, A168.

——, and Coauthors, 2016: Volcanic air pollution over the Island of Hawai'i: Emissions, dispersal, and composition. Association with respiratory symptoms and lung function in Hawai'i Island school children. *Environ. Int.*, **92–93**, 543–552, https://doi.org/10.1016/j.envint.2016.03.025.

Tofte, K., P. Chu, and G. M. Barnes, 2017: Large-scale weather patterns favorable for volcanic smog occurrences on O'ahu, Hawai'i. *Air Qual. Atmos. Health*, **10**, 1163–1180, https://doi.org/10.1007/s11869-017-0502-z.

Unger, D. A., 1985: A method to estimate the continuous ranked probability score. Preprints, *Ninth Conf. on Probability and Statistics in Atmospheric Sciences*, Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

USGS, 2019a: Volcano updates archive. Accessed 23 August 2019, https://volcanoes.usgs.gov/vhp/archive_search.htm.

——, 2019b: Frequently asked questions about Mauna Loa volcano. Accessed 23 August 2019, https://volcanoes.usgs.gov/observatories/hvo/faq_maunaloa.html.

Weigel, A. P., and S. J. Mason, 2011: The generalized discrimination score for ensemble forecasts. *Mon. Wea. Rev.*, **139**, 3069–3074, https://doi.org/10.1175/MWR-D-10-05069.1.

Wilks, D. S., and T. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, https://doi.org/10.1175/MWR3402.1.

World Meteorological Organization, 2012: *Guidelines on Ensemble Prediction Systems and Forecasting*. WMO Tech. 1091, World Meteorological Organization, 23 pp.

Worth, R. M., 1995: Respiratory impacts associated with chronic VOG exposure on the Island of Hawai'i. *Proc. Vog and Laze Seminar*, Honolulu, HI, Hawaii State Department of Health.