Atmos. Chem. Phys., 18, 4019–4038, 2018
https://doi.org/10.5194/acp-18-4019-2018

Atmospheric Chemistry and Physics
Open Access
EGU

Volcanic ash modeling with the NMMB-MONARCH-ASH model: quantification of offline modeling errors

Alejandro Marti and Arnau Folch

Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain

Correspondence: Alejandro Marti (alejandro.marti@bsc.es)

Received: 18 April 2017 – Discussion started: 23 May 2017
Revised: 19 October 2017 – Accepted: 21 November 2017 – Published: 22 March 2018

Abstract. Volcanic ash modeling systems are used to simulate the atmospheric dispersion of volcanic ash and to generate forecasts that quantify the impacts from volcanic eruptions on infrastructures, air quality, aviation, and climate. The efficiency of response and mitigation actions is directly associated with the accuracy of the volcanic ash cloud detection and modeling systems. Operational forecasts build on offline coupled modeling systems in which meteorological variables are updated at the specified coupling intervals. Despite the concerns from other communities regarding the accuracy of this strategy, the quantification of the systematic errors and shortcomings associated with the offline modeling systems has received no attention. This paper employs the NMMB-MONARCH-ASH model to quantify these errors by employing different quantitative and categorical evaluation scores. The skills of the offline coupling strategy are compared against those from an online forecast considered to be the best estimate of the true outcome. Case studies are considered for a synthetic eruption with constant eruption source parameters and for two historical events, which suitably illustrate the severe aviation disruptive effects of European (2010 Eyjafjallajökull) and South American (2011 Cordón Caulle) volcanic eruptions. Evaluation scores indicate that systematic errors due to the offline modeling are of the same order of magnitude as those associated with the source term uncertainties. In particular, traditional offline forecasts employed in operational model setups can result in significant uncertainties, failing to reproduce, in the worst cases, up to 45–70 % of the ash cloud of an online forecast. These inconsistencies are anticipated to be even more relevant in scenarios in which the meteorological conditions change rapidly in time. The outcome of this paper encourages operational groups responsible for real-time advisories for aviation to consider employing computationally efficient online dispersal models.

1 Introduction

Volcanic ash modeling systems are used to simulate the atmospheric dispersion of volcanic ash and to generate operational short-term forecasts to support civil aviation and emergency management. These systems are vital in efforts to prevent aircraft flying into ash clouds, which could result in catastrophic impacts (e.g., Miller and Casadevall, 2000; Prata and Tupper, 2009). The aviation community is concerned about the detection and tracking of volcanic ash clouds to provide timely warnings to aircrafts and airports. In the event of an eruption, the individual Volcanic Ash Advisory Center (VAAC) responsible for the affected region combines ash cloud satellite observations and dispersal simulations to issue periodic volcanic ash advisories (VAAs). These are text and graphical products informing on the extent of the ash clouds at relevant flight levels and their forecasted trajectories at 6, 12, and 18 h ahead that are updated periodically or whenever significant changes occur in the eruption source term. All this information is used to ensure flight safety by supporting critical decisions such as closure of ash-contaminated air space and airports or diversion of aircraft flight paths to prevent encounters. The noteworthy economic impact and social disruption of these air traffic restrictions are, therefore, directly associated with the accuracy of the volcanic ash cloud detection and modeling systems.

Volcanic ash modeling systems require (i) a source term to characterize the emission of ash, (ii) a meteorological model (MetM) for the description of the atmospheric con-

ditions, and (iii) a volcanic ash transport and dispersal model (VATDM; e.g., Folch, 2012) to forecast the particle transport and deposition mechanisms. The MetM and the VATDM can be coupled either online or offline. In an offline modeling system, the MetM runs a priori and independently from the VATDM to produce the required meteorological fields at regular time intervals, e.g., every 1 to 6 h for typical mesoscale and global operational MetM outputs, respectively. Meteorological fields are then furnished to the VATDM, which commonly assumes constant values for these fields during each time coupling interval or, at most, performs a linear interpolation in time. This approach is convenient in terms of computing time because different VATDM model executions are possible without rerunning the meteorological component, e.g., to update the source term whenever the eruption conditions vary, for inverse modeling of ash emissions (e.g., Marti et al., 2016; Webster et al., 2012), or to perform an ensemble forecast (e.g., Galmarini et al., 2010) in which all the ensemble members share the same meteorological conditions. However, offline MetM and VATDM coupling introduces model and numerical errors due to non-synchronized time stepping, use of unaligned grids and projections, and/or inconsistencies in the numerical schemes. In contrast, in an online modeling system, the MetM and the VATDM run concurrently and consistently and the particle transport is automatically tied to the model resolution time and space scales, resulting in a more realistic representation (e.g., Grell and Baklanov, 2011; Marti et al., 2017). At present, all operational volcanic ash forecast systems follow the offline approach and the few existing online atmospheric chemistry and transport models specific for volcanic ash, e.g., WRF-Chem (Stuefer et al., 2013) or NMMB-MONARCH-ASH (Marti et al., 2017) are still restricted to a research level. However, notwithstanding the increase in computational power in recent years, the experiences from other fields (e.g., online models for air quality, dust), and the fact that the total computing time required to run an online coupled model is actually not substantially larger (e.g., Bowman et al., 2013; Grell and Baklanov, 2011; Marti et al., 2016), the benefits of the traditional offline systems are at question.

Since the 2010 Eyjafjallajökull eruption in Iceland, considerable effort and progress has been made to quantify and reduce ash cloud modeling and forecasting errors associated with a number of critical aspects (e.g., Bonadonna et al., 2012, 2015a) including, among others, characterization of the source term and related uncertainties in model inputs (e.g., Costa et al., 2016), model parameterization of relevant physical phenomena (e.g., aggregation, particle settling velocities, deposition mechanisms), propagation of errors in the driving MetM forecast, or satellite detection and retrieval algorithms. However, and surprisingly, the quantification of shortcomings associated with the offline coupling strategy has received no attention, despite the fact that lessons from other communities show that these can be substantial (e.g., Baklanov et al., 2014). Errors implicit in offline coupled systems include inaccurate handling of atmospheric processes occurring on timescales smaller than the coupling interval and eventual feedbacks between the volcanic ash cloud and meteorology. These inconsistencies are anticipated to be more relevant in scenarios in which the meteorological conditions (mainly wind speed and direction) change rapidly in time and/or for long-range transport simulations model-coupling errors in time.

The objective of this paper is to quantify the model shortcomings and systematic errors associated with traditional offline forecasts. In that context, we employ the strategies available in the NMMB-MONARCH-ASH model to evaluate the predictability of the offline coupling approach against that from an online forecast considered to be the best estimate of the true outcome. Section 2 describes the methodology used to quantify the coupling model errors; Sect. 3 presents the results from a synthetic case study with constant eruption source parameters (ESPs) and the systematic errors attributed to the meteorological coupling intervals. Section 4 evaluates the results from two real cases that illustrate disruptive effects of European (2010 Eyjafjallajökull) and South American (2011 Cordón Caulle) eruptions. Section 5 discusses the magnitude of the model forecast errors implicit in the offline approach by comparing it with better-constrained errors, e.g., in eruption source parameters. Finally, Sect. 6 provides the conclusions of this work.

2 Methods

2.1 Modeling background

NMMB-MONARCH-ASH (Marti et al., 2017) is a novel online meteorological and atmospheric transport model to simulate the emission, transport, and deposition of tephra (ash) and aerosol particles released during a volcanic eruption. The model predicts ash cloud trajectories, concentration of ash at relevant flight levels, and the expected ground deposit for both regional and global domains. The online coupling in NMMB-MONARCH-ASH allows for solving both meteorology and tephra–aerosol transport concurrently and interactively at every time step. The model attempts to pioneer the forecast of volcanic ash and aerosols by embedding a series of new modules on the Barcelona Supercomputing Center (BSC) operational system for short-term and mid-term chemical weather forecasts (NMMB-MONARCH, formerly known as NMMB/BSC-CTM; Badia et al., 2017; Jorba et al., 2012) developed at the BSC in collaboration with the US National Centers for Environmental Prediction (NCEP) and the NASA Goddard Institute for Space Studies. Its meteorological core, the Nonhydrostatic Multiscale Model on the B-grid (NMMB – Janjic and Gall, 2012), is a fully compressible meteorological model with a non-hydrostatic option that allows for nested global–regional atmospheric simulations

by using consistent physics and dynamics formulations. The NMMB model became the North American Mesoscale Forecast System (NAM) operational meteorological model in October of 2011, and it has been computationally robust, efficient, and reliable in operational applications and preoperational tests since then. In high-resolution numerical weather prediction applications, the efficiency of the model significantly exceeds that of several established state-of-the-art non-hydrostatic models (e.g., Janjic and Gall, 2012). The computational efficiency of its meteorological core suggests that NMMB-MONARCH-ASH could be used in an operational setting to forecast volcanic ash (Marti et al., 2017).

The model allows for two different coupling strategies: online and offline. In the online version of the model the MetM and VATDM are fully integrated in one unified modeling system, using one main time step for integration (e.g., consistent spatial and temporal interpolation, map projections, dataset inputs, numerical schemes), which is updated at each MetM model time step (e.g., 30 s). This coupling strategy offers a more realistic representation of the meteorological conditions, improving the current state of volcanic ash dispersal models, especially in situations in which meteorological conditions change rapidly in time, two-way feedbacks are significant, or distal ash cloud dispersal simulations are required. In contrast, in the offline version, the model uses "effective wind fields" in which meteorological conditions (e.g., wind velocity, mid-layer pressure) are set to constant and are only updated at the user-defined coupling interval. This strategy aims to replicate the decoupling effect of traditional VATDM dispersal models used at the operational level. However, note that most operational offline systems perform a linear interpolation in time, thereby attenuating the offline coupling effects. This is not possible in our offline strategy because of the concurrent solution of both meteorology and dispersal.

2.2 Forecasts

The skills of an atmospheric dispersal model are known to vary in space and time. In that context, NMMB-MONARCH-ASH simulations are performed to account for the effects of the coupling interval and the dispersal distance of the forecast. Online forecasts are evaluated against simulations from four different offline coupling intervals (i.e., 1, 3, 6, and 12 h) to compare the skills of each offline coupling approach. For this purpose, model comparisons are performed for (i) a synthetic case study with constant ESPs to focus exclusively on the effect of the offline coupling interval and (ii) two historical cases accounting for the effects of changing the ESPs, including a case in which meteorological conditions change rapidly in time (first phase of the 2011 Cordón Caulle eruption) and a case in which these changes are less abrupt (first phase of the 2010 Eyjafjallajökull eruption). Finally, in order to assess the order of magnitude of the error associated with the offline forecasts, we compare it with the better-constrained source of forecast error attributed to the source term (i.e., uncertainties in column height and related mass eruption rate), known to be one of the main reasons (first order) for VATDM output variability (e.g., Bonadonna et al., 2010).

Forecasts (offline and online) for each application use the same computational domain and share the same spatial and temporal scales, allowing for a gridded (point-to-point) evaluation. The standard NMMB-MONARCH-ASH parameterization is employed for all simulations (Marti et al., 2017). The meteorological driver is initialized with wind fields from the ERA-Interim reanalysis at $0.75° \times 0.75°$ resolution, and for regional domains the reanalysis also furnishes 6 h boundary conditions. For the purpose of this study, forecasts predict ash cloud trajectories and concentration of ash at relevant flight levels for a period of up to 48 h. This approach is consistent with most volcanic ash forecast operational systems.

2.3 Evaluation methods

In general terms, forecast evaluation is the process of assessing the goodness of a model prediction. The forecast is compared, or verified, against a corresponding observation of what actually occurred or some good estimate of the true outcome. For the purpose of this work, the output from the online forecast is considered to be the model "observations" (i.e., best estimate of the true outcome) and is compared with those results from the different offline forecasts. However, it is important to highlight that the aim of these simulations is not to reconstruct the actual eruptive events but to compare the skills of the offline forecasts with the online forecasts in order to quantify their differences.

The accuracy of a volcanic ash forecast can be measured by means of different evaluation scores as no single evaluation score is adequate to fully determine the goodness of a VATDM forecast. Consequently, a detailed assessment of the strengths and weaknesses of a set of forecasts normally requires more than one or two scores (Jolliffe and Stephenson, 2012). In this study, we evaluate the skills of the offline versus online NMMB-MONARCH-ASH forecasts in terms of their ash column loading (ACL) using different quantitative and categorical evaluation scores. These scores are often based on grid points; they compare observations and predictions per grid cell and compute various metrics for the entire set or subset of grid points. Objects from both online and offline ACL fields must be identified for each evaluation score. An object is a group of adjacent grid cells that have an ACL value above a given threshold. Here, the threshold is defined as the typical ash detection limit for most satellite retrievals ($\sim 0.2\,\mathrm{g\,m^{-2}}$; Prata and Prata, 2012). Modeled ACL values below this threshold are omitted from all evaluation metrics.

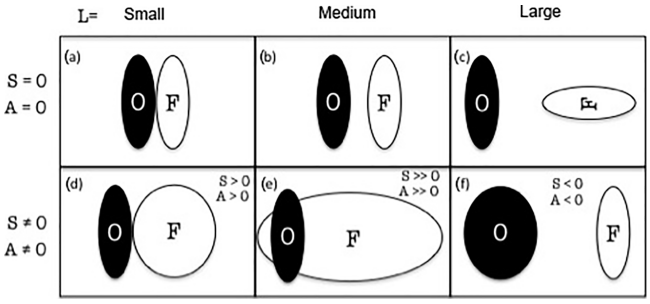

Figure 1. Schematic representation of the possible online (O, representing the "observations") and offline forecast (F) combinations of the different components for the quantitative object-based metric SAL: structure (S), amplitude (A), and location (L). Modified from Wernli et al. (2008).

2.3.1 Quantitative evaluation scores

Quantitative evaluation scores are useful to determine the degree to which a forecast differs from the best estimate of the true outcome (i.e., the online simulation). Quantitative measures such as correlation coefficients, RMSE, or bias are simple in implementation and are thus regularly used to compare and monitor the quality of a forecast. Here, we use RMSE to assess the average magnitude of forecast errors, bias to assesses the difference between the online and offline forecast means, and the Pearson's correlation coefficient to reflect the linear association of the forecasts. Due to their invariance properties, these measures are considered to be suitable in many predictive sciences, and in particular in weather and climate forecasting (Jolliffe and Stephenson, 2012).

However, the skill of a dispersion forecast is known to vary in space and time, making these commonly used evaluation scores problematic for grid-point-based measures. A classical example to illustrate these limitations is the "double penalty problem" (Wernli et al., 2008), in which a forecast is correct in terms of amplitude, size, and timing, but slightly incorrect concerning location, resulting, for example, in very poorly rated correlation and RMSE scores. To overcome these limitations, we complement the previous scores with the quantitative object-based metric SAL (Wernli et al., 2008). This metric individually considers aspects of the structure (S), amplitude (A), and location (L) of a forecast, revealing meaningful information about the systematic differences between forecasts. This diagnostic metric has been previously used to measure the skill of volcanic ash forecast using data insertion from satellite observations (Wilkins et al., 2016) and has been adapted here to compare the quality of online and offline coupled NMMB-MONARCH-ASH forecasts. Figure 1 provides a schematic representation of different metric combinations and scores in SAL.

The S (Eq. 1) component in SAL captures information about the size and shape of ACL objects (Wilkins et al., 2016) by computing the normalized weighted mean mass difference (V, Eq. 2) for the online and offline forecasts:

$$S = \frac{V_{\text{off}} - V_{\text{on}}}{0.5\,|V_{\text{off}} + V_{\text{on}}|}. \tag{1}$$

Weighted means (V) of the ash load fields in the column are estimated considering the total mass (R_n) and the scaled mass (V_n, Eq. 3) for the number of objects in the domain (M):

$$V = \frac{\sum_{n=1}^{M} R_n V_n}{\sum_{n=1}^{M} R_n}. \tag{2}$$

Scaled masses (V_n) for all objects are calculated separately for each object as follows:

$$V_n = \sum_{(xy)\in O_n} R_{xy}/R_n^{\max}, \tag{3}$$

where xy is the grid cell location within the forecasted field, R_{xy} is the area-integrated concentration field (i.e., ash column mass in grams per square meter) in grid cell xy, and $R_n^{\max}$ is the maximum grid cell ash mass in object O_n. Note that, in the case of a single object, $V = V_n$. Structure scores range between $[-2, 2]$, with positive values indicating more objects in the offline forecast and that ACL values are too spread out and/or flat. A negative S score occurs when the offline forecast ACL objects cover too small of an area or are too peaked (or a combination of both).

The A component corresponds to the normalized difference of the domain-average ash mass values (R). This provides a simple measure of the quantitative accuracy of the total concentration of ash in the domain, ignoring the field's subregional structure:

$$A = \frac{\bar{R}_{\text{off}} - \bar{R}_{\text{on}}}{0.5\,\left|\bar{R}_{\text{off}} + \bar{R}_{\text{on}}\right|}, \tag{4}$$

where $\bar{R}_{\text{off}}$ and $\bar{R}_{\text{on}}$ are the ash masses averaged over all grid cells in the domain, i.e., $\bar{R}_n = \sum_{(xy)\in\text{Domain}} R_{xy} \big/ \text{domain area}$. Amplitude scores range between $[-2, 2]$, with 0 denoting no difference between offline and online forecasts. An amplitude score of $+1$ or -1 indicates that offline forecasts overestimate or underestimate, respectively, the domain-averaged ACL by a factor of 3. Scores of $A = 0.4$ and 0.67 correspond to factors of 1.5 and 2, respectively (Wernli et al., 2008).

The L component of SAL compares the mass distribution between forecasts. The L component is composed of two parts:

$$L = L_1 + L_2. \tag{5}$$

The first one (L_1, Eq. 6) compares the normalized distance between the center of mass (C) of the offline and online ACL

fields over the maximum distance within the entire domain (d):

$$L_1 = \frac{|C_{\text{off}} + C_{\text{on}}|}{d}. \quad (6)$$

The values of L_1 are in the range of $[0, 1]$, with $L_1 = 0$ suggesting identical centers of mass for both forecasts. However, separated ash clouds could also have the same center of mass, and therefore $L_1 = 0$ would not necessarily indicate a perfect match. The second part of the location component (L_2, Eq. 7) aims to distinguish such situations by measuring the weighted average (H, Eq. 8) between the center of mass of the total ash load and the center of mass for each object (C_n):

$$L_2 = 2\left[\frac{|H_{\text{off}} + H_{\text{on}}|}{d}\right] \quad (7)$$

$$H = \frac{\sum_{n=1}^{M} R_n \, |C_{\text{off}} + C_{\text{on}}|}{\sum_{n=1}^{M} R_n}. \quad (8)$$

In the event that both online and offline ACL fields have only one object, then $L_2 = 0$. A factor of 2 is used to scale L_2 to the range of L_1. Hence, the total location component of L can reach values between $[0, 2]$ and can only be possible for an offline forecast in which both the distance between objects and the center of mass agree with the online forecast. It is important to mention that since both offline and online computational domains are the same, the magnitude dependency of L on the size of the domain does not affect our interpretation of this SAL component.

Absolute SAL scores range from 0 to 6, with scores closest to 0 denoting the best agreement between forecasts. The computation of the structure and location components of SAL groups accounts for objects (i.e., adjacent grid cells) with a value above a given threshold for the forecasted variable. For this study, objects are given as $O_n, n = 1, \ldots, M$, where M is the number of objects in the model domain. Each object combines at least two grid cells to avoid unrealistic single ash-containing grid cells. As defined previously, the object identification threshold for the ACL is set to $0.2\,\text{g}\,\text{m}^{-2}$. Modeled ACL values below this threshold are omitted from all components of SAL.

2.3.2 Categorical evaluation scores

From an operational perspective, it is also important to know whether the presence of volcanic ash constitutes an airspace threat or not. In that context, the significance of quantitative volcanic ash forecasts can be measured in terms of categorical evaluation scores (Jolliffe and Stephenson, 2012). These scores are less sensitive to larger errors than quantitative evaluation scores. This is particularly important for extremely skewed data such as ACL, providing the degree to which the forecast supports a decision maker during an emergency event (i.e., closure of airspace). Consequently, ash loads can be viewed categorically (or binary for "yes" or "no" events) according to whether that value exceeds a threshold (event) or not (non-event). Here, we compute a series of categorical evaluation scores based on a contingency table (Table 1), which describes the combined distribution of forecast events and non-events for each coupling strategy. In Table 1, "hits" represents the number of grid points for which both forecasts (offline and online) exceed the threshold previously established ($0.2\,\text{g}\,\text{m}^{-2}$); "misses" represents the number of points for which only online forecasts exceed this threshold; "false alarms" indicates the number of points for which only offline forecasts exceeded the threshold; and finally, "correct negatives" represents the number of points for which neither offline nor online forecasts exceeded the threshold value. In this paper, we use these binary skill metrics to calculate four categorical evaluation scores:

a. Probability of detection (POD) measures the fraction of ash points observed in the online forecast and that were correctly predicted for the offline forecast. This score is good for rare events, should be used together with the FAR score (FAR, Eq. 10), and is insensitive to false alarms. The POD score can reach values between [0,1]:

$$\text{POD} = \frac{\text{hits}}{(\text{hits} + \text{misses})}; [0, 1]. \quad (9)$$

b. False alarm ratio (FAR) measures the fraction of ash points predicted by the offline forecast that were observed to be non-events (i.e., not exceeding the threshold) in the online forecast. This score should be used together with the previous POD score and ignores the misses. The FAR score can reach values between [0, 1]:

$$\text{FAR} = \frac{\text{false alarm}}{(\text{hits} + \text{false alarm})}; [0, 1]. \quad (10)$$

c. Frequency bias (FBI) measures the ratio of frequency of offline forecast points to the frequency of observed ash points in the online forecast. This score indicates whether the forecast system has a tendency to underestimate (FBI < 1) or overestimate (FBI > 1) events. However, it does not measure how well the offline forecast corresponds to the online simulation and only measures relative frequencies. The FBI score can reach values between $[0, \infty]$:

$$\text{FBI} = \frac{(\text{hits} + \text{false alarm})}{(\text{hits} + \text{misses})}; [0, \infty]. \quad (11)$$

d. Critical success rate (CSI) measures the fraction of all offline and online forecast points that were correctly diagnosed and does consider both misses and false alarms. The CSI score can reach values between [0, 1]:

$$\text{CSI} = \frac{\text{hits}}{(\text{hits} + \text{misses} + \text{false alarms})}; [0, 1]. \quad (12)$$

Table 1. Contingency table of binary events for categorical verification scores at each grid point.

Offline forecast exceeding threshold	Online forecast exceeding threshold	
	Yes	No
Yes	Hits	False alarms
No	Misses	Correct negatives

Similar metrics, such as the Figure of Merit in Space (FMS; Galmarini et al., 2010), have been used in previous works (e.g., Wilkins et al., 2016) to complement the SAL score for the evaluation of the spatial coverage between forecasts:

$$\mathrm{FMS} = \frac{B_{\mathrm{off}} \cap B_{\mathrm{on}}}{B_{\mathrm{off}} \cup B_{\mathrm{on}}}\,;\,[0,1], \tag{13}$$

where B_{off} and B_{on} are the modeled and observed ACL areas, respectively. In both cases, a score of 1 suggests a complete spatial overlap of the evaluated forecast. Alternatively, the spatial overlap will decrease as these scores reach values close to 0. Here, we employ the FMS metric to evaluate the spatial coverage of the forecasts and to complement a missing spatial coverage component in SAL. To be consistent with our implementation of SAL, the spatial ash coverage is computed only for forecast ACL fields exceeding a threshold of $0.2\,\mathrm{g\,m^{-2}}$. However, it is worth mentioning that a low FMS score could also suggest two similar shapes shifted in space (Mosca et al., 1998) and, therefore, should be used together with the SAL score.

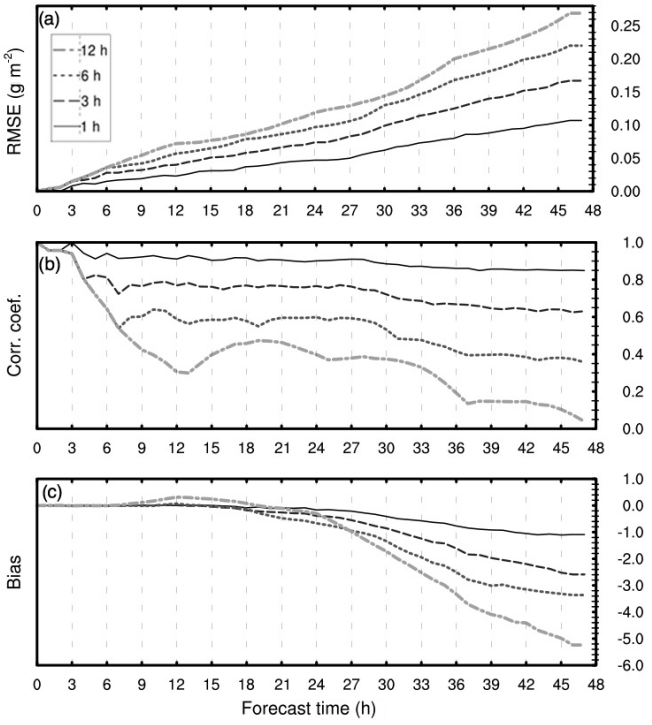

Figure 2. Quantitative evaluation scores for NMMB-MONARCH-ASH synthetic application: **(a)** root mean square error; **(b)** Pearson's correlation coefficient; **(c)** error bias.

3 Synthetic case study

The first step of our evaluation consists of isolating the model's shortcomings and systematic errors that are exclusively associated with the offline coupling strategy employed in traditional volcanic ash forecasts. To this purpose, we constructed a preliminary synthetic case based on the first 48 h of the 2011 Caulle eruption with constant ESPs. This synthetic application reduces the differences associated with the source term (i.e., different source term quantification because of different wind fields) and allows us to isolate the systematic errors coming from the offline coupling approach. Within this framework, we limited the eruption duration to 12 h, using a constant column height and employing the Mastin et al. (2009) relationship (mass eruption rate vs. column height) for the dispersion evaluation of a single bin of ash (one particle class) during the first 48 h of the event. Multiple regional simulations of NMMB-MONARCH-ASH were performed to produce four different offline coupled forecasts, in which meteorological variables are updated at the specified coupling intervals (i.e., 1, 3, 6, and 12 h). Details about the 2011 Caulle eruption, accompanying meteorological conditions and the computational domain, are described in Sect. 4.2. Table 2 provides a summary of the ESPs used for this application. The skills of these forecasts were compared against the online coupled simulation, employing the quantitative and categorical evaluation scores described in Sect. 2.3. Scores at the end of the simulation (48 h) are shown in Table 3. For the purpose of this paper, we focus on describing the scores for the 3 and 6 h offline coupled forecasts, representative of an operational forecast driven by a global MetM.

Figure 2 shows the results of the quantitative evaluation scores, RMSE (Fig. 2a), correlation coefficient (Fig. 2b), and bias (Fig. 2c), as a function of the forecast's length for each coupling interval in the synthetic case. These scores assist in determining the degree to which offline forecasts correspond to the best estimate of the true outcome (online forecast). In general terms, and as expected a priori, all scores

indicate that the quality of the forecast decreases with longer coupling intervals (i.e., 1, 3, 6, and 12 h) and length of the forecast. The RMSE score is presented in Fig. 2a and is used to assess the average magnitude of the offline forecast errors. Figure 2b shows how the linear association between the online and offline forecasts (Pearson's correlation coefficient) significantly decreases with longer coupling intervals, reaching noticeably low correlations. For example, for a 3 h coupled (i.e., meteorological data coupled in intervals of every 3 h) forecast the resulting coefficient after 24 h of simulation is below 0.6, while for a 6 h coupled forecast (coupled in intervals of every 6 h) it is below 0.5, and below 0.4 after 48 h. These scores indicate that traditional offline forecasts are not capable of reproducing more than half of the true outcome, suggesting that the coupling frequency in tephra dispersal modeling could be a critical source of error. This result is relevant considering that most emergency model setups employ 3 h temporal resolution forecasts with update frequencies of 6 h or more (Bowman et al., 2013; WMO, 2012). Finally, Fig. 2c depicts the forecast bias over that from the online simulation. In general terms, all offline forecasts underestimate ACL, reaching values between -1 and $-5\,\mathrm{g\,m^2}$ at the end of the forecast for the 1 and 12 h coupling intervals, respectively (Table 3).

Figure 3 illustrates the results from the quantitative object-based metric SAL, aimed at evaluating the variation in space and time of the forecasts. As with previous scores, the error associated with the SAL score also increases with the length of the coupling frequency. For all offline simulations within the synthetic case, the structure component of the metric (Fig. 3a) explains most of the discrepancy with the online forecast. Negative values of S indicate that offline forecasts predict fields that cover too small of an area and/or are too peaked. In addition, results from the amplitude and location components indicate a slight underestimation of the domain-averaged ACL for most offline forecasts, employing comparable centers of mass with the online reference. In general terms, systematic differences in the offline forecast are 3 and 5 times higher for a coupling frequency of 3 and 6 h than those of the 1 h interval (Table 3).

Categorical scores resulting from the evaluation of the synthetic case are summarized in Fig. 4. As in the previous scores, a threshold value of $0.2\,\mathrm{g\,m^{-2}}$ is considered to define the ash-contaminated objects, categorizing these as yes or no events depending on if they exceed the threshold or not. These scores are critical for the aviation industry during a volcanic crisis since they can determine the closure of the airspace or the cancellation of flights. Figure 4a illustrates the POD for each forecast. As expected, this metric clearly shows how the probability of detecting ash-contaminated points in the offline forecasts decreases with longer coupling intervals and the forecast length. In addition, this figure also suggests that POD scores decrease considerably during the first hours of the forecast, matching the time for which the source was active. This trend is applicable to all categorical

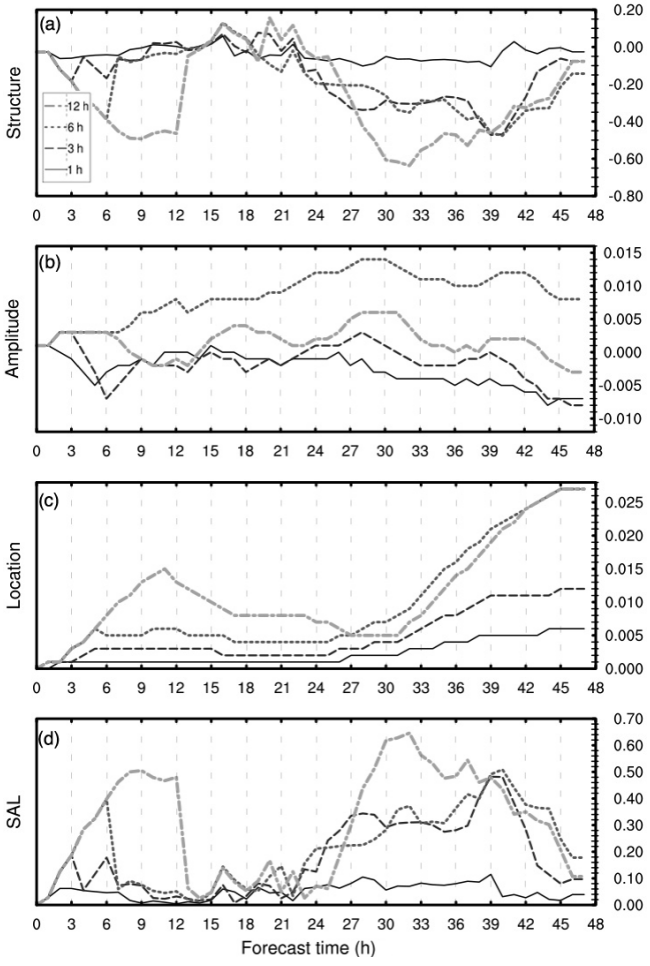

Figure 3. SAL evaluation scores for NMMB-MONARCH-ASH synthetic case: **(a)** structure; **(b)** amplitude; **(c)** location; **(d)** combined SAL.

evaluation scores. After 48 h, the POD scores for the 3 and 6 h coupled forecasts are 0.752 and 0.603, respectively. For coupling frequencies above 6 h, the probability of detecting ash-contaminated areas drops below 50 % (e.g., 12 h coupled forecast in Table 3). POD scores are complemented by the results from the FAR metric, which measures the fraction of ash events predicted by the offline forecasts that were observed to be non-events. FAR scores (Fig. 4b) are consistent with POD scores, predicting 25 % of false ash-contaminated objects after 48 h of simulation (6 h coupled forecast). The equivalent plot for the FBI metric as a function of forecast length is shown in Fig. 4c. This metric indicates that all forecasts tend to underestimate the ACL, especially while the eruption is active. After that time, FBI scores stabilize between values ranging from 0.7 to 1.0. Finally, Fig. 4d illustrates the spatial overlap between offline and online forecasts defined by the FMS. This metric provides similar results to the POD metric. However, in this case, false alarms (Table 1) are also considered in the metric, leading to FMS

Table 2. Summary of eruption source parameters (ESPs) used in NMMB-MONARCH-ASH for the synthetic case and the 2010 Eyjafjallajökull and 2011 Cordón Caulle applications.

Source term	Synthetic	2010 Eyjafjallajökull	2011 Cordón Caulle
Run duration	12 h	96 h	72 h
Vertical distribution of mass in the column	Suzuki (1983) distribution ($A = 4$, $L = 5$)	Suzuki (1983) distribution ($A = 4$, $L = 5$)	Suzuki (1983) distribution ($A = 4$, $L = 5$)
Mass eruption rate (MER)	Mastin et al. (2009)	Degruyter and Bonadonna (2012) (Fig. 5b)	Degruyter and Bonadonna (2012) (Fig. 10a)
Column height (above vent)	8500 m	Fig. 5c	Fig. 10b
TGSD	1 bin ($\Phi = 6$; $\rho = 1933\,\mathrm{kg\,m^{-3}}$; sphericity $= 0.9$)	Bonadonna et al. (2011)	Bonadonna et al. (2015b)
Sedimentation model	Ganser (1993)	Ganser (1993)	Ganser (1993)

Table 3. Evaluation scores for the synthetic case at the end of the 48 h forecast with NMMB-MONARCH-ASH. Metrics: Pearson's correlation coefficient (R); root mean square error (RMSE); error bias (BiAS); structure (S); amplitude (A); location (L); absolute SAL (| SAL |); probability of detection (POD); false alarm ratio (FAR); frequency bias (FBI); and Figure of Merit in Space (FMS).

Coupling/score	R	RMSE	BIAS	S	A	L	| SAL |	POD	FAR	FBI	FMS
1 h	0.849	0.107	−1.090	−0.026	−0.007	0.006	0.039	0.897	0.039	0.934	0.855
3 h	0.631	0.167	−2.589	−0.077	−0.008	0.012	0.097	0.752	0.108	0.843	0.669
6 h	0.357	0.220	−3.362	−0.143	0.008	0.027	0.178	0.603	0.243	0.796	0.477
12 h	0.039	0.269	−5.239	−0.077	−0.003	0.027	0.107	0.400	0.413	0.682	0.281

scores lower than those for the POD metric. Considering this, FMS scores indicate that the spatial overlap (i.e., probability of hits over hits, misses, and false alarms) between the online and the 3 h coupled offline forecast after 48 h of simulation is around 65 % and below 50 % for the 6 h coupled forecast (Table 3).

4 Application examples

In addition to the synthetic case, we present two applications of NMMB-MONARCH-ASH for the simulation of the initial phases of the 2010 Eyjafjallajökull and 2011 Cordón Caulle eruptions. In these cases, offline forecasts are evaluated taking into account the effects of the coupling interval and the actual changes in the ESPs (i.e., mass eruption rate (MER) depending on wind field) for each event. A summary of the ESPs used for each application is presented in Table 2. These two events have shed light on the importance of ash dispersal in the context of aviation safety (Bonadonna et al., 2012), and they suitably illustrate the severe disruptive effects of European and South American eruptions. Similar to the synthetic case, online and offline forecasts were compared on the same temporal scales and spatial grid; a gridded (point-to-point) evaluation was performed between forecasts following the criteria presented in the contingency Table 1; the output of the online forecast was considered as the "observed" (best estimation of true outcome) field; and a threshold of $0.2\,\mathrm{g\,m^{-2}}$ was employed as the ACL detection limit.

For each application we include (i) a brief description of the eruptive event; (ii) a summary of the modeling setup to simulate the eruption; and (iii) a comprehensive evaluation of the plume dispersal forecast including qualitative, quantitative, and categorical evaluations and metrics. For the purpose of summarizing the results of these evaluation scores, we focus on describing those scores from the 6 h coupled forecast.

4.1 The 2010 Eyjafjallajökull eruption

The April 2010 eruption of Eyjafjallajökull volcano (63.6° N, 19.6° W, vent height 1666 m a.s.l.) in southern Iceland created unprecedented disruptions to European air traffic during 15–20 April. On 14 April a major outbreak of the central crater under the covering ice cap led to surface activity, causing phreatomagmatic explosions, generation of volcanic ash, and eruption columns rising up to 9 km (a.s.l) (Institute of Earth Sciences, 2010). The initial ash clouds traveled rapidly across the North Atlantic and North Sea, reaching southern Norway on 15 April and then traveling southwards as a frontal cloud crossing over to many northern European countries. In turn, the London VAAC dispatched immediate warnings to European aviation authorities and other VAAC centers every 3–6 h. The southern part of the ash cloud finally grounded in the northern part of the Alps. On 20 April

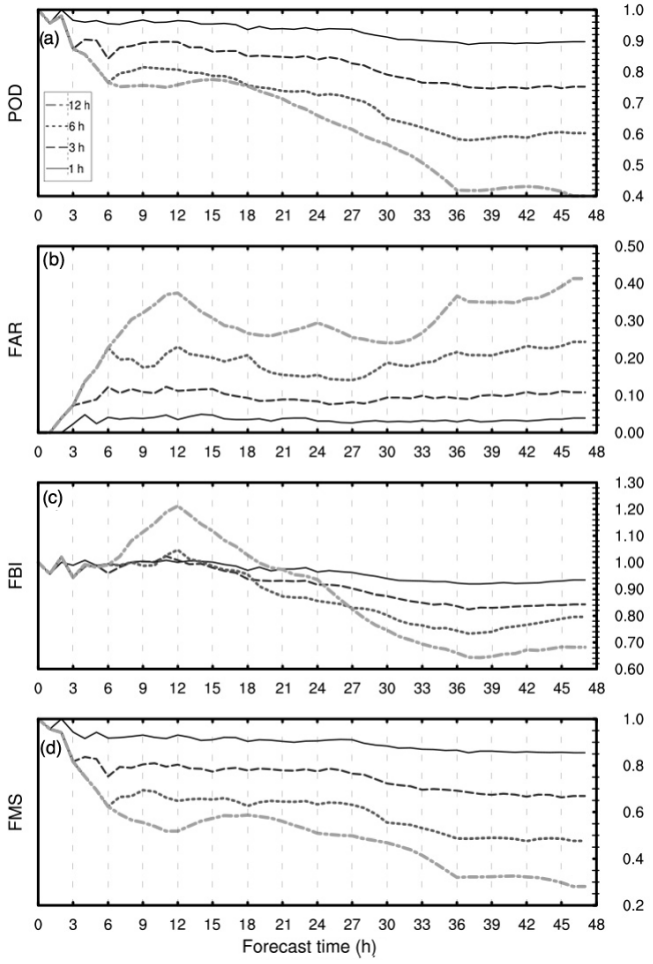

Figure 4. Categorical evaluation scores for the NMMB-MONARCH-ASH synthetic case including **(a)** probability of detection (POD), **(b)** false alarm ratio (FAR), **(c)** frequency bias (FBI), and **(d)** Figure of Merit in Space (FMS).

new guidelines based on safe ash concentration thresholds were adopted, allowing for the ability to resume operations in large areas previously banned. In addition, several other ash cloud episodes occurred during late April and May, disrupting the European airspace for a total of 13 days (over 4 million passengers stranded due to cancellation or delay of over 100 000 flights), affecting 25 countries, and costing the aviation industry billions of euro (Oxford Economics, 2010). These impacts brought into focus how significantly volcanoes can affect communities and economies far away from the source and the critical importance of accurate volcanic ash forecasts.

4.1.1 Modeling setup

For the purpose of simulating this eruption, we employed NMMB-MONARCH-ASH with a model domain consisting of 401×428 grid points, covering the northern and western regions of Europe and using a grid with a horizontal resolution of $0.15° \times 0.15°$ and 60 vertical layers. The top pressure of the model was set to 10 hPa (~ 26 km) with a mesh refinement near the top (to capture the dispersion of ash) and the ground (to capture the characteristics of the atmospheric boundary layer). The computational domain spans in longitude from 30° W to 30° E and in latitude from 34° S to 84° N. The ESPs characterizing the event are described in Table 2 and presented in Fig. 5. Figure 5a shows the variations in column height for the duration of the forecast. Figure 5b illustrates the results from estimating the MER using the different formulations available in NMMB-MONARCH-ASH (Marti et al., 2017). Figure 5c shows the MER variations associated with the different offline coupling intervals with the MetM and compare them with the fully coupled online forecast.

4.1.2 Qualitative evaluation

Figure 6 shows the plume dispersal (ash column loading; ACL) from the online forecast corresponding to the first explosive phase (14–18 April) of the Eyjafjallajökull eruption (Gudmundsson et al., 2012). This phase is conveniently divided into 14–16 April, when the volcanic plume produced a well-defined sector towards the east, and 17 to early 18 April, when northerly winds drove the plume to the south. Complementary to this figure, Fig. 7 illustrates the airspace contamination forecasted by the model during the first phase of the eruption at flight levels FL050 and FL200. This figure illustrates the ash hazard aviation guidelines, which distinguish zones of low (green, ash concentration less than $0.2\,\mathrm{mg\,m^{-3}}$), moderate (orange, ash concentration between 0.2 and $2\,\mathrm{mg\,m^{-3}}$), and high (red, ash concentration above $2\,\mathrm{mg\,m^{-3}}$) concentration of ash employed to regulate no-fly zones. This information is critical for air traffic management to assist flight dispatchers while planning flight paths and designing alternative routes in the presence of a volcanic eruption. Model results show the volcanic cloud traveling E-NE, achieving critical concentration values in northern Europe during 15–17 June, and suggesting severe disruptions in the European airspace.

Figure 8 shows a qualitative comparison between the online and the different offline coupled forecasts for Eyjafjallajökull application. Qualitative comparisons are presented for each coupling interval in different rows (i.e., first row: 1 h; second row: 3 h; third row: 6 h; fourth row: 12 h coupling). Areas in grey (hits) represent grid points for which both forecasts (offline and online) exceed the established threshold. Red areas (misses) indicate those regions where the offline forecast fails to predict existing ash (underprediction). Finally, blue areas (false alarms) illustrate those domain areas for which only offline forecasts exceed the threshold, implying a false prediction of ash (overprediction). In general terms, offline forecasts for the Eyjafjallajökull event tend to overpredict towards the north of the plume and to underpredict towards the south. While results of the 1 h offline fore-

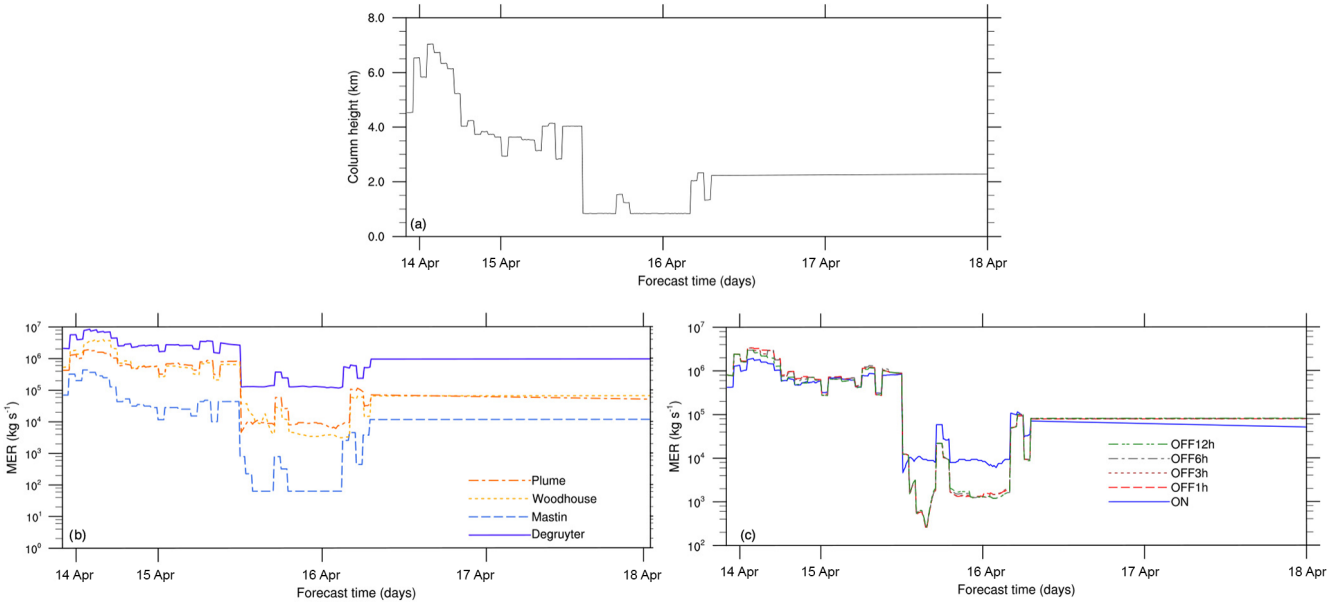

Figure 5. Eruption source parameters for the 2010 Eyjafjallajökull application: **(a)** column height fluctuation over time; **(b)** resulting MER over time considering different plume parameterizations (FPLUME – Folch et al., 2016; Woodhouse – Woodhouse et al., 2013; Mastin – Mastin et al., 2009; Degruyter – Degruyter and Bonadonna, 2012); **(c)** resulting MER for each coupling strategy (meteorology coupled online or with intervals of time of 1, 3, 6, and 12 h) with the Degruyter option only.

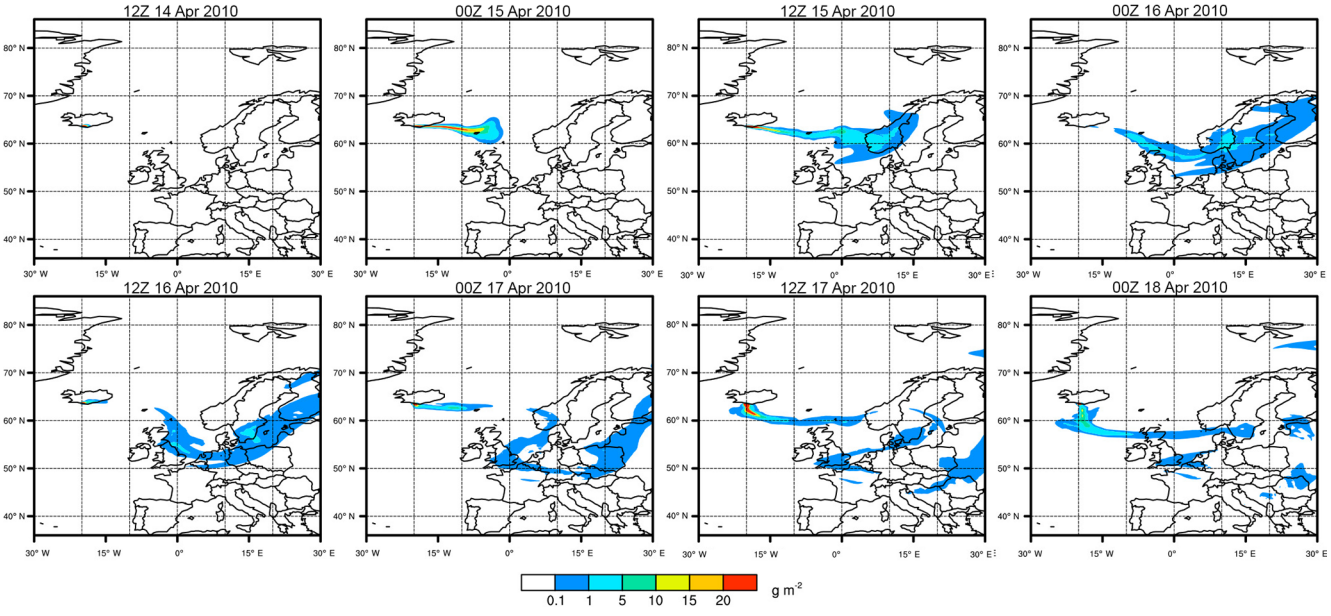

Figure 6. NMMB-MONARCH-ASH total ACL (mass loading; g m^{-2}) for the 2010 Eyjafjallajökull application.

cast indicate mostly hits (H), Fig. 8 clearly suggests that the number of missed (M) and false alarm (FA) points increases with longer coupling intervals and the length of the forecast. This is due to the prevailing mid-troposphere south-oriented wind acceleration over the south of Iceland and the North Sea (Folch et al., 2012), and it is consistent with those results presented previously in the synthetic case. As a consequence, these forecasts would miss, for example, the arrival of volcanic ash over northern Germany in the late afternoon of 16 April as indicated by the Deutscher Wetterdienst (DWD) ceilometer network at the time of the eruption (Flentje et al., 2010). As a general approximation, Fig. 8 suggests that for the Eyjafjallajökull eruption, offline forecasts with

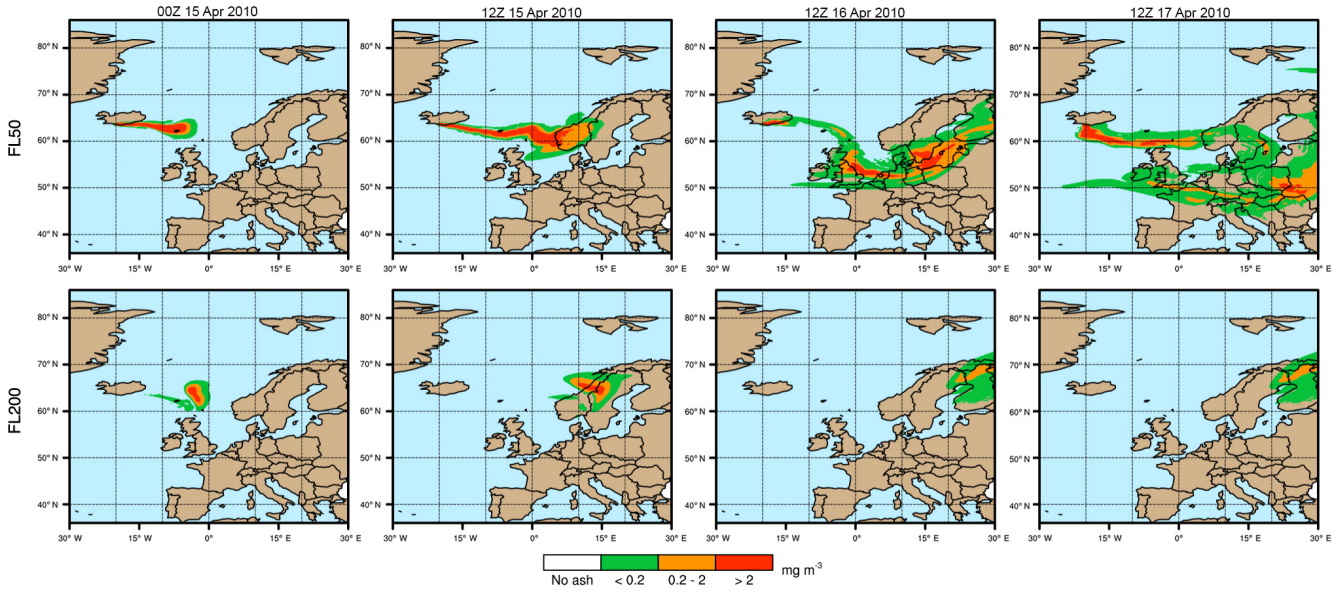

Figure 7. Ash hazard aviation guidelines applied for 2010 Eyjafjallajökull application over time. Zones of low (green, ash concentration $< 0.2\,\mathrm{mg\,m^{-3}}$), moderate (orange, ash concentration between 0.2 and $2\,\mathrm{mg\,m^{-3}}$), and high (red, ash concentration above $2\,\mathrm{mg\,m^{-3}}$) concentration are displayed for FL50 (top) and FL200 (bottom).

coupling intervals of 3 h and above could result in significant inconsistent predictions (M + FA areas).

4.1.3 Quantitative and categorical evaluation

Figure 9 shows the results of the quantitative and categorical metrics for the ACL offline forecasts for the 2010 Eyjafjallajökull application. Complementing this figure, Table 4 shows the scores for all coupled forecasts after 48 h from the eruption starting time. As found in the synthetic case, quantitative and categorical metrics lessen their scores for longer coupling intervals and forecast lengths.

Quantitative evaluation scores RMSE (Fig. 9a), correlation coefficient (Fig. 9b), and bias (Fig. 9c) show more comparable trends than those reported for the synthetic case. After 48 h of simulation, the 3 h coupled forecast scores show barely any correlation with the online forecast and a RMSE of $0.122\,\mathrm{g\,m^{-2}}$. Bias scores suggest that all offline forecasts tend to underestimate ACL between -0.33 and $-2.5\,\mathrm{g\,m^{2}}$ at the end of the forecast. Figure 9d–g illustrate the results from the quantitative object-based metric SAL, quantifying the variation in space and time of the forecasts. For the Eyjafjallajökull application, both the structure (Fig. 9d) and amplitude (Fig. 9e) from the offline forecasts explain most of the discrepancy with the online forecast. Contrary to the synthetic case, in which the A component had a residual effect towards the total SAL, in this application the offline forecasts tend to underestimate the amplitude component by 20 % for coupling intervals equal to or above 3 h. This is anticipated since meteorological conditions are updated at each given interval and no additional ash-contaminated objects are found in the domain. After the first coupling with the MetM takes place, scores start to stabilize. After this time, the A component stabilizes, reducing its associated underestimation factor. Location scores (Fig. 9f) suggest a comparable mass distribution of the ACL fields for the online and offline forecasts. Finally, absolute SAL scores after 48 h of simulations (Table 4) indicate that systematic differences in the offline forecast are approximately 2 times higher for a coupling frequency of 3 h than those of the 1 h interval.

Categorical scores for the Eyjafjallajökull application are summarized in Fig. 9h–k. Results from the POD metric (Fig. 9h) show that the probability of detecting ash-contaminated events in the offline forecasts decreases with longer coupling intervals, especially during the time the first coupling with the MetM occurs. After 48 h, the POD score for the 3 h coupled forecast is around 65 % and below 50 % (i.e., 0.46) for the 6 h coupled forecast. Conversely, results from the FAR metric follow an increasing trend (Fig. 9i), misrepresenting nearly 45 % objects in the domain. Results from the FBI metric (Fig. 9j) indicate that all offline forecasts tend to underestimate the ACL. Finally, FMS scores suggest that the spatial overlap between the online and the offline forecasts after 48 h of simulation is below 50 % for those simulations with coupling intervals of 3 h or more.

4.2 The 2011 Cordón Caulle eruption

The 2011 Cordón Caulle eruption exemplifies a typical mid-latitude Central and South Andean eruption. The Cordón Caulle volcanic complex (Chile, 40.5° S, 72.2° W, vent height 1420 m a.s.l.) reawakened on 4 June 2011 around

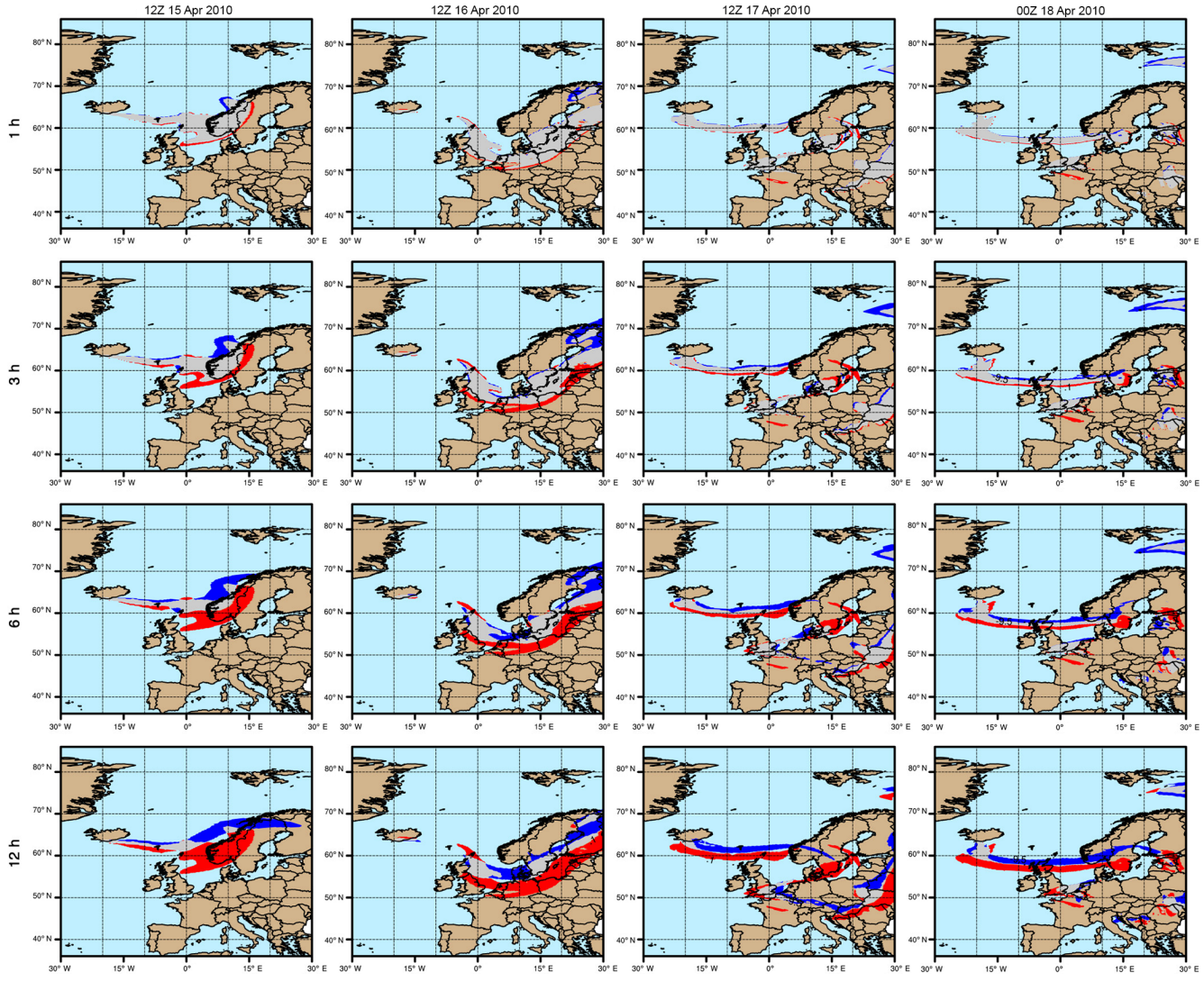

Figure 8. Qualitative comparison between the online and offline forecasts with 1 h (row 1), 3 h (row 2), 6 h (row 3), and 12 h (row 4) coupling intervals. Gridded evaluation is performed following the criteria presented in the contingency Table 1. Hit (grey), missed (red), and false alarm (blue) predictions are shown for the 2010 Eyjafjallajökull case over time.

Table 4. Evaluation scores for the 2010 Eyjafjallajökull eruption application at the end of the 48 h forecast with NMMB-MONARCH-ASH. Metrics: Pearson's correlation coefficient (R), root mean square error (RMSE), error bias (BiAS), structure (S), amplitude (A), location (L), absolute SAL (|SAL|), probability of detection (POD), false alarm ratio (FAR), frequency bias (FBI), and Figure of Merit in Space (FMS).

Coupling/score	R	RMSE	BIAS	S	A	L	|SAL|	POD	FAR	FBI	FMS
1 h	0.787	0.072	−0.327	−0.087	−0.019	0.006	0.112	0.881	0.091	0.969	0.805
3 h	0.386	0.122	−0.783	−0.138	−0.049	0.019	0.206	0.664	0.283	0.926	0.499
6 h	0.08	0.149	−1.82	−0.116	−0.063	0.034	0.213	0.465	0.438	0.828	0.332
12 h	−0.292	0.177	−2.521	−0.136	−0.105	0.051	0.292	0.251	0.671	0.762	0.156

18:30 UTC after decades of quiescence. The initial explosive phase spanned over more than 2 weeks, generating ash clouds that dispersed over the Andes (Collini et al., 2013). The climatic phase (∼ 27 h) (Jay et al., 2014) was associated with a ∼ 9 km (a.s.l.) high column (Osores et al., 2014). For the period between 4 and 14 June, numerous flights and airports were disrupted in Paraguay, Uruguay, Chile, southern Argentina, and Brazil (Wilson et al., 2013).

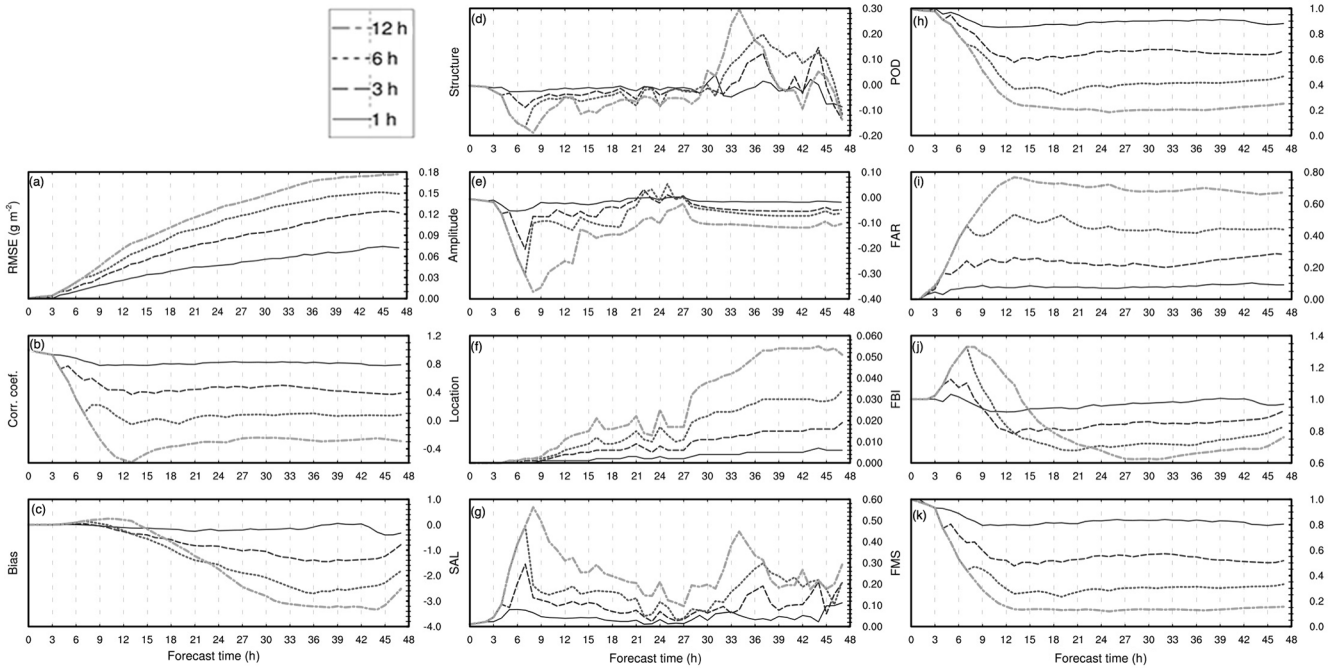

Figure 9. Online vs. offline evaluation scores for the 2010 Eyjafjallajökull case.

4.2.1 Modeling setup

The model domain for this application consists of 268×268 grid points covering the northern regions of Chile and Argentina using a horizontal resolution of $0.15° \times 0.15°$ and 60 vertical layers. The top pressure of the model was set to 10 hPa ($\sim$ 26 km). The computational domain spans in longitude from 41 to 81° W and in latitude from 18 to 58° S. The ESPs characterizing the Caulle event are described in Table 2. Figure 10a shows the slight variation in column height for the duration of the forecast (Osores et al., 2014). Figure 10b illustrates the results from simulating the MER over time considering different coupling strategies.

4.2.2 Qualitative evaluation

Figure 11 illustrates the plume dispersion from the online forecast associated with the early Plinian phase (4–7 June) of the Cordón Caulle eruption. The initial ash cloud reached the Atlantic coast on 4 June late afternoon (Collini et al., 2013), just before turning to the northeast to reach the northern part of Argentina during the 6 June and the city of Buenos Aires in the days after. The effect of the plume dispersion on air-traffic management is shown in Fig. 12. This figure shows the airspace contamination forecasted by the model during 4–6 June at flight levels FL050 and FL200. Model results show the volcanic cloud achieving critical concentration values within a wide area east of the Andes range. On 6 June, simulation results show the volcanic cloud moving east, threatening the main international airports that service the province of Buenos Aires. These results suggest that the cancellation of multiple flights in several Argentinean airports during this time was justified.

Figure 13 shows the qualitative comparison between the online and offline coupled forecasts for the first days of the 2011 Cordón Caulle eruption. In this case, given that the plume height during the first hours of the eruption was more consistent (no significant changes in wind speed and direction) than for the Eyjafjallajökull application, the difference between forecasts is less suggestive, although still remarkable. Contrary to the Eyjafjallajökull application, offline forecasts tended to underestimate to the north of the plume and slightly overestimate to the south. The resulting evaluation from these inconsistencies indicates that offline forecasting with longer coupling intervals missed the abrupt shift in the plume course known to be associated with early 6 June. This alteration was due to the change in the wind direction toward the N-NE first and then again towards the SE (e.g., Elissondo et al., 2016). As a consequence, these results suggest that offline forecasts would miss the correct arrival time of volcanic ash to the main international airports in Buenos Aires a few days later (i.e., Ezeiza and Aeroparque Jorge Newbery airports).

4.2.3 Quantitative and categorical evaluation

Figure 14 summarizes the results for the quantitative and categorical metrics for the 2011 Caulle application. Metric scores at the end of the simulation are presented in Table 5. Results for the Cordón Caulle application are consistent with

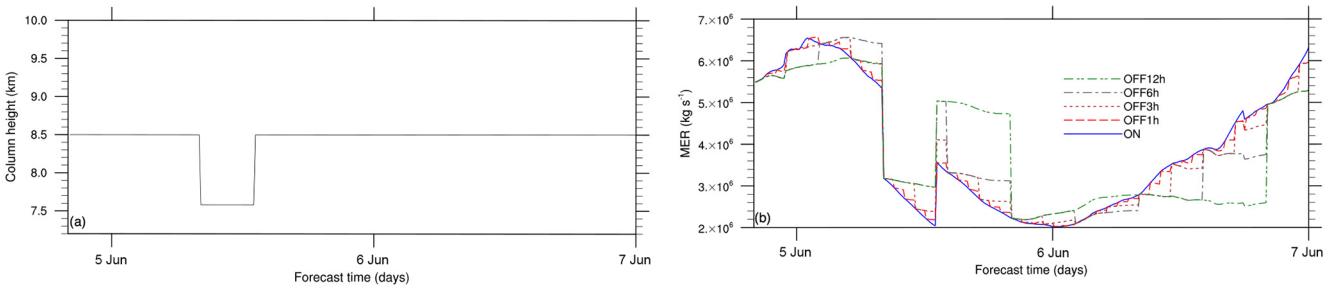

Figure 10. Eruption source parameters for the 2011 Cordón Caulle case: **(a)** column height fluctuation over time. **(b)** Resulting MER over time for each coupling strategy (meteorology coupled online or with intervals of time of 1, 3, 6, and 12 h).

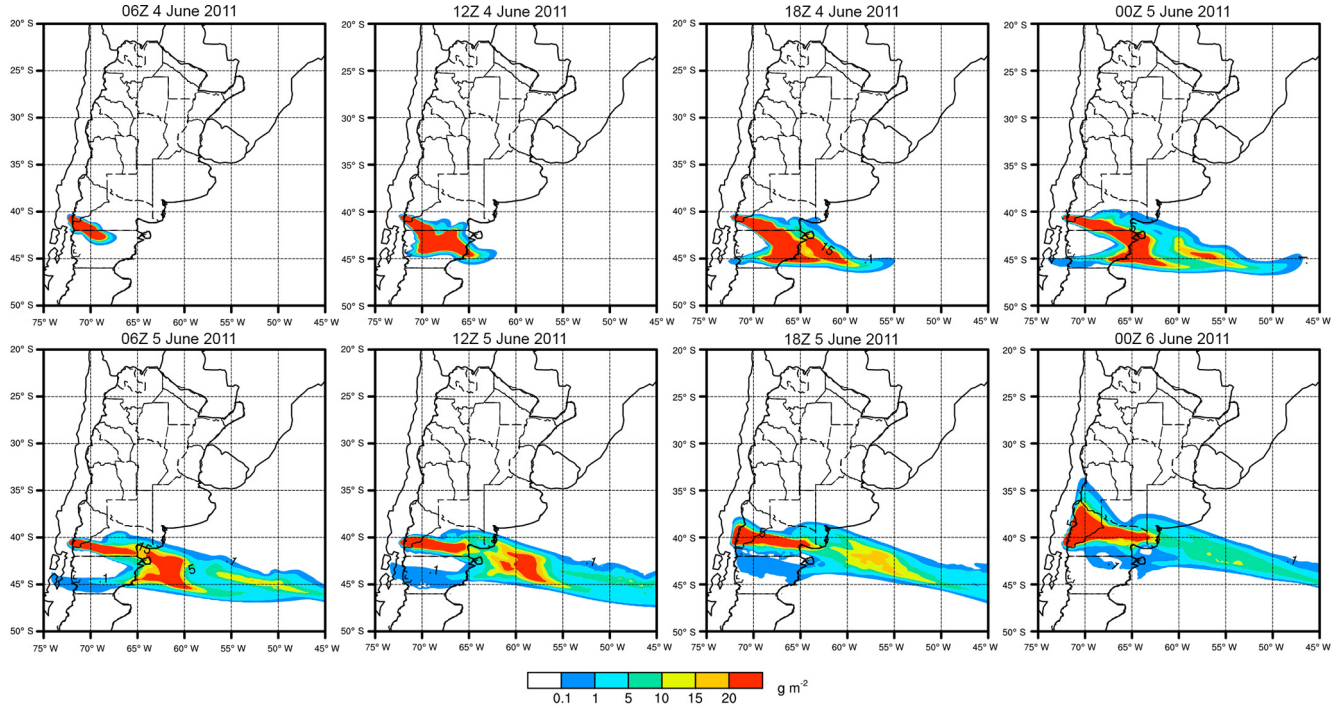

Figure 11. NMMB-MONARCH-ASH total column load (mass loading; g m^{-2}) for the 2011 Cordón Caulle case.

those from the synthetic case and the Eyjafjallajökull application in that the uncertainty of the forecast increases significantly with the length of the coupling frequency employed. Quantitative evaluation scores RMSE (Fig. 14a), correlation coefficient (Fig. 14b), and bias (Fig. 14c) show comparable trends to those from the Eyjafjallajökull application. Despite this similarity, linear correlation coefficients between offline and online forecasts for the Caulle application are higher than those from the Eyjafjallajökull simulation. This result is explained by the fewer changes in the source term (e.g., variations in the column height) during the Caulle event. After 48 h of simulation, the 3 h coupled forecast scores a correlation coefficient of 0.80 with a RMSE of 0.11 g m^{-2}, while the 6 h coupled forecast scores 0.6 and 0.16 g m^{-2}, respectively. Bias scores suggest that all offline forecasts tend to underestimate ACL between -0.13 and -4.75 g m^{-2} at the end of the forecast.

Figure 14d–g illustrate the results from the quantitative object-based metric SAL for the Cordón Caulle event. SAL scores (Fig. 14g) suggest that differences between online and offline strategies for the Cordón Caulle application are considerably higher than those for the Eyjafjallajökull application. This is due to the changing meteorological conditions during to the Cordón Caulle event and confirms that inconsistences associated with offline forecasts are more relevant in scenarios in which the meteorological conditions (mainly wind speed and direction) vary rapidly in time. In terms of the individual components of SAL, structure (Fig. 14d) and amplitude (Fig. 14e) scores explain most of the discrepancy with the online forecast. Structure scores indicate that more objects occur in the offline forecast and ACL values are too

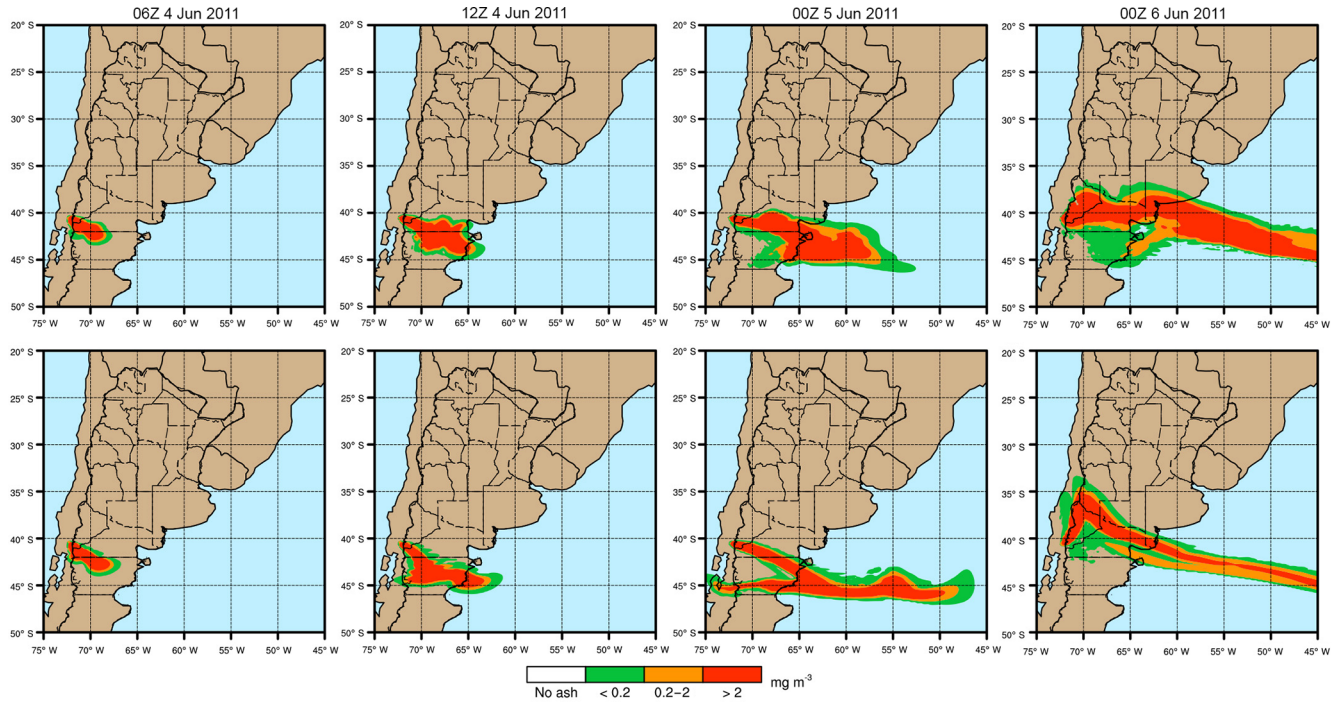

Figure 12. Ash hazard aviation guidelines applied for the 2011 Cordón Caulle application over time. Zones of low (green, ash concentration $< 0.2\,\mathrm{mg\,m^{-3}}$), moderate (orange, ash concentration between 0.2 and $2\,\mathrm{mg\,m^{-3}}$), and high (red, ash concentration above $2\,\mathrm{mg\,m^{-3}}$) concentration are displayed for FL50 (top) and FL200 (bottom).

Table 5. Evaluation scores for the 2011 Cordón Caulle eruption application at the end of the 48 h forecast with NMMB-MONARCH-ASH. Metrics: Pearson's correlation coefficient (R), root mean square error (RMSE), error bias (BiAS), structure (S), amplitude (A), location (L), absolute SAL (|SAL|), probability of detection (POD), false alarm ratio (FAR), frequency bias (FBI), and Figure of Merit in Space (FMS).

Coupling/score	R	RMSE	BIAS	S	A	L	|SAL|	POD	FAR	FBI	FMS
1 h	0.932	0.068	−0.131	−0.001	−0.02	0.002	0.023	0.965	0.028	0.993	0.934
3 h	0.808	0.113	−1.16	−0.008	−0.038	0.008	0.054	0.881	0.063	0.94	0.82
6 h	0.598	0.164	−3.657	0.123	−0.105	0.017	0.245	0.719	0.114	0.811	0.639
12 h	0.333	0.212	−4.754	0.033	−0.287	0.046	0.366	0.568	0.248	0.755	0.453

spread out and/or flat, while amplitude scores suggest that offline forecasts tend to underestimate the total concentration of ash in the domain up to 50 %. The systematic error associated with the offline forecasts is clearly demonstrated after 18 h of simulation (Fig. 14g), time during which changes in wind speed and direction start to be noteworthy (Fig. 11). Location scores (Fig. 14f) suggest a consistent mass distribution of the ACL fields amongst forecasts. The Cordón Caulle application is a perfect example to illustrate the importance of complementing traditional quantitative metrics with the quantitative object-based metric SAL. For this particular case, SAL scores are able to capture the inconsistences of the offline dispersion forecast due to the changing meteorological conditions that other quantitative metrics (i.e., RMSE, correlation coefficient, bias) cannot account for.

Finally, categorical scores for the Cordón Caulle application are presented in Fig. 14h–k. Results suggest that the skill of the forecast decreases with longer coupling intervals, following trends similar to those found in the Eyjafjallajökull application. After 48 h, FMS scores suggest that the spatial overlap between the online and the 6 h coupled offline forecasts is below 65 % and around 80 % for the 3 h coupled forecast (Table 5; Fig. 14k), with a probability of misrepresenting ash-contaminated objects by around 10 % (FAR; Fig. 14i). Results from the FBI metric (Fig. 14j) indicate that all offline forecasts tend to underestimate the ACL.

5 Discussion

Volcanic ash modeling systems are used to simulate the atmospheric dispersion of volcanic ash and to generate forecasts to quantify the impacts from volcanic eruptions on air quality, aviation, and climate. However, volcanic ash

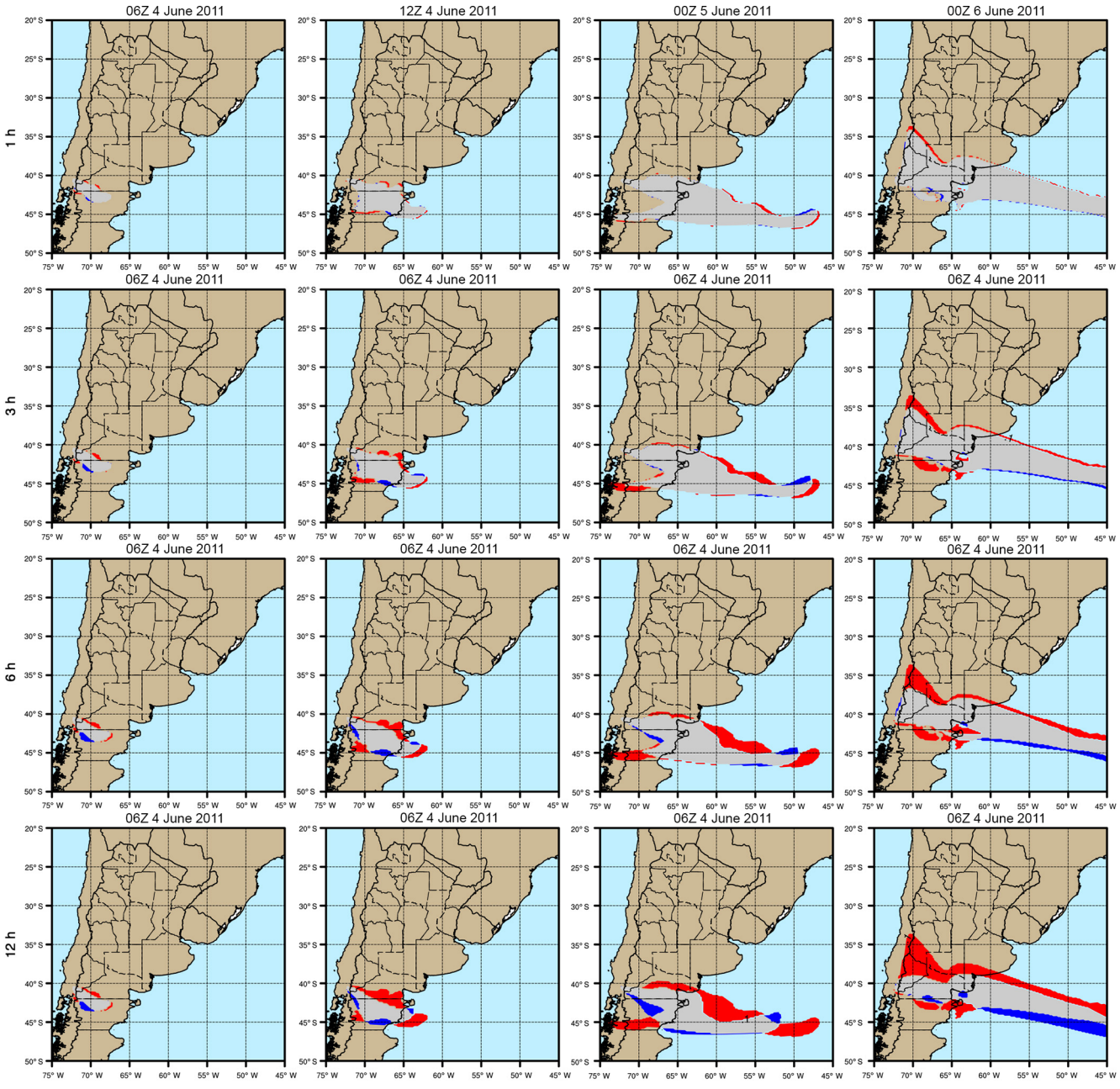

Figure 13. Qualitative offline vs. online forecast comparison for 1 h (row 1), 3 h (row 2), 6 h (row 3), and 12 h (row 4) coupling intervals. Gridded evaluation is performed following the criteria presented in the contingency Table 2. Hit (grey), missed (red), and false alarm (blue) predictions are shown for the 2011 Cordón Caulle application over time.

forecasts require the consideration of numerous and complex uncertainties. The 2010 Eyjafjallajökull eruption clearly demonstrated the need for a better understanding of the uncertainties associated with the dispersal model employed in operational volcanic ash forecasting. Since then, the scientific community has focused on identifying and improving uncertainties primarily associated with the characterization of the source term (e.g., MER, column height). However, and surprisingly, the quantification of systematic errors and shortcomings associated with the meteorological data driving the dispersion model has received little attention. Traditionally, operational volcanic ash forecasts employ offline coupling strategies to produce the required meteorological fields at regular time intervals, e.g., every 1 or 6 h for typical mesoscale and global operational MetM configurations, respectively. In previous sections of this paper, we have shown the meaningful negative impact of employing offline coupling intervals on the accuracy of the ash-cloud simulations

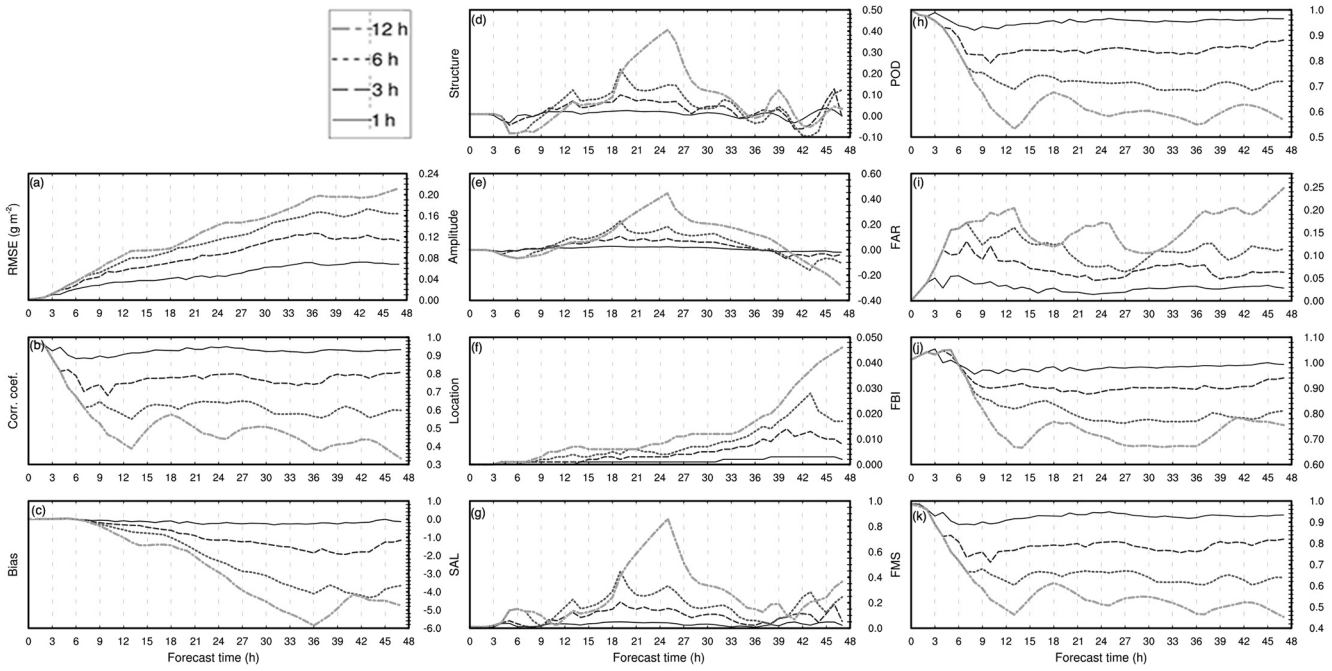

Figure 14. Online vs. offline evaluation scores for the 2011 Cordón Caulle application.

as compared to online coupled forecasts. In particular, Sect. 3 showed the scores from evaluating a synthetic case focusing exclusively on the effect of the coupling approach. Evaluation scores reveal that the uncertainty of the offline forecasts increases significantly with the length at which the meteorological driver is coupled with the dispersion model (e.g., up to 4 times for the 6 h coupled forecast). However, the question of how this compares to other better-constrained sources of forecast error still remains unanswered.

This section aims to answer this question by evaluating to what extent the magnitude of the model forecast errors implicit in the offline approach compares with that of the source term. For this purpose, we performed four additional experimental simulations under the synthetic case in which ESP for the online forecast were modified by (i) employing a ± 2 MER factor (i.e., $\times 2$ and $1/2$ times the original MER) and (ii) varying the corresponding column height by $\pm 20\,\%$. As in previous applications, experimental forecasts were evaluated on the same temporal scales and spatial grid against the online forecast employing a range of complementary quantitative and categorical metrics. Figure 15 illustrates the evaluation scores from these four simulations and compares them with those of the 6 h coupled offline forecast. Overall, Fig. 15 reveals that systematic errors and shortcomings associated with the traditional offline coupling strategies employed in operational volcanic ash forecast are of the same order of magnitude as those uncertainties attributed to the characterization of the source term. For example, correlation coefficients (Fig. 15b) and POD scores (Fig. 15h) suggest an additional 10–30 % level of uncertainty attributed to the 6 h coupled forecast than those associated with the source term. In that same context, at the end of the simulation, FMS scores (Fig. 15k) reveal that the spatial overlap between the online and the 6 h coupled offline forecasts is $\sim 20\,\%$ lower than that from varying the column height by $\pm 20\,\%$, and $\sim 50\,\%$ lower than those from altering the original MER.

These results suppose a significant advance in the quantification of the uncertainty sources associated with traditional offline volcanic ash forecasts.

6 Conclusions

This paper quantifies the systematic errors inherent in offline coupling strategies employed for operational volcanic ash forecasting. For this purpose, we employed the different coupling strategies available in the NMMB-MONARCH-ASH model to evaluate the predictability limitations of the offline forecast against the online forecast. Model comparisons were performed for a synthetic case study focusing exclusively on the effect of the coupling approach, and for two historical cases accounting for changing meteorological conditions and ESPs. Evaluation scores indicate that systematic errors credited by offline forecast are of the same order of magnitude as those better-constrained uncertainties associated with the source term. In particular, offline forecasts in operational setups can result in significant errors in the dispersion of the ash plume for coupling intervals above 3 h. The results of this study show that 3 h coupling offline forecasts fail to reproduce about 35 % of the online forecast (best estimate of the true outcome) for a case with constant ESPs (synthetic

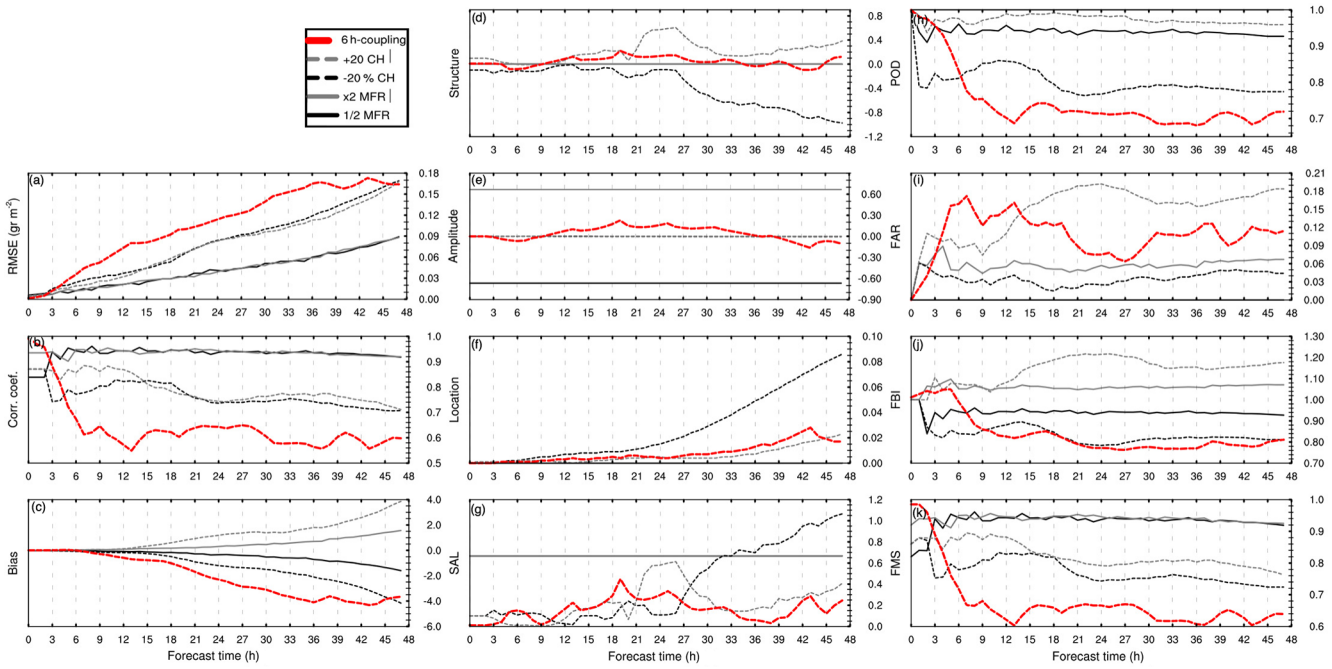

Figure 15. Online vs. offline evaluation scores for the NMMB-MONARCH-ASH synthetic application representing the uncertainty associated with the source term. ESPs were modified for the eruption column height (±20 %) and MER (×2 and 1/2). Scores are compared with those from the 6 h offline coupled forecasts (red line).

case): close to 50 % for the 2010 Eyjafjallajökull case and 20 % for the 2011 Cordón Caulle case. For offline forecasts with coupling intervals above 3 h, these discrepancies are significantly higher. These inconsistencies are anticipated to be even more relevant in scenarios in which the meteorological conditions change rapidly in time. The outcome of this paper advocates that operational groups responsible for real-time advisories for aviation consider employing computationally efficient online dispersal models.

Data availability. All datasets used are free to access from a third party. The ECMWF reanalysis dataset ERA-Interim is available at http://apps.ecmwf.int/datasets/ (European Centre for Medium-Range Weather Forecasts).

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under the project NEMOH (REA grant agreement no. 289976). We are extremely grateful to Oriol Jorba for providing critical insights of the NMMB-MONARCH model, and to Maria Soledad Osores and Costanza Bonadonna for sharing data to validate this work. We also thank Zavisa Janjic, Matthew Hort, and Larry Mastin for providing constructive comments, which helped to improve the clarity and the quality of the paper. Meteorological data were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). Numerical simulations were performed at the Barcelona Supercomputing Center with the MareNostrum III Supercomputer using 512 – 8×4 GB DDR3-1600 DIMMS (2GB/core) Intel Sandy Bridge processors, iDataPlex Compute Racks, a Linux operating system, and an InfiniBand interconnection.

Edited by: Stefano Galmarini
Reviewed by: Larry Mastin and one anonymous referee

References

Badia, A., Jorba, O., Voulgarakis, A., Dabdub, D., Pérez García-Pando, C., Hilboll, A., Gonçalves, M., and Janjic, Z.: Description and evaluation of the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (NMMB-MONARCH) version 1.0: gas-phase chemistry at global scale, Geosci. Model Dev., 10, 609–638, https://doi.org/10.5194/gmd-10-609-2017, 2017.

Baklanov, A., Schlünzen, K., Suppan, P., Baldasano, J., Brunner, D., Aksoyoglu, S., Carmichael, G., Douros, J., Flemming, J., Forkel, R., Galmarini, S., Gauss, M., Grell, G., Hirtl, M., Joffre, S., Jorba, O., Kaas, E., Kaasik, M., Kallos, G., Kong, X., Korsholm, U., Kurganskiy, A., Kushta, J., Lohmann, U., Mahura, A., Manders-Groot, A., Maurizi, A., Moussiopoulos, N., Rao, S. T., Savage, N., Seigneur, C., Sokhi, R. S., Solazzo, E., Solomos, S., Sørensen, B., Tsegas, G., Vignati, E., Vogel, B., and Zhang, Y.: Online coupled regional meteorology chemistry models in Europe: current status and prospects, Atmos. Chem. Phys., 14, 317-398, https://doi.org/10.5194/acp-14-317-2014, 2014.

Bonadonna, C., Folch, A., Loughlin, S., and Puempel, H.: Report on the IAVCEI-WMO workshop on Ash Dispersal Forecast and Civil Aviation, IUGG Newsletter, 11, 6–7, 2010.

Bonadonna, C., Genco, R., Gouhier, M., Pistolesi, M., Cioni, R., Alfano, F., Hoskuldsson, A., and Ripepe, M.: Tephra sedimentation during the 2010 Eyjafjallajkull eruption (Iceland) from deposit, radar, and satellite observations, J. Geophys. Res.-Sol. Ea., 116, B12202, https://doi.org/10.1029/2011JB008462, 2011.

Bonadonna, C., Folch, A., Loughlin, S., and Puempel, H.: Future developments in modelling and monitoring of volcanic ash clouds: Outcomes from the first IAVCEI-WMO workshop on Ash Dispersal Forecast and Civil Aviation, B. Volcanol., 74, 1–10, https://doi.org/10.1007/s00445-011-0508-6, 2012.

Bonadonna, C., Biass, S., and Costa, A.: Physical characterization of explosive volcanic eruptions based on tephra deposits: Propagation of uncertainties and sensitivity analysis, J. Volcanol. Geoth. Res., 296, 80–100, https://doi.org/10.1016/j.jvolgeores.2015.03.009, 2015a.

Bonadonna, C., Cioni, R., and Pistolesi, M.: Sedimentation of long-lasting wind-affected volcanic plumes?: the example of the 2011 rhyolitic Cordón Caulle eruption, Chile, B. Volcanol., 77, 13–32, https://doi.org/10.1007/s00445-015-0900-8, 2015b.

Bowman, K. P., Lin, J. C., Stohl, A., Draxler, R., Konopka, P., Andrews, A. and Brunner, D.: Input data requirements for Lagrangian trajectory models, B. Am. Meteorol. Soc., 94, 1051–1058, https://doi.org/10.1175/BAMS-D-12-00076.1, 2013.

Collini, E., Osores, M. S., Folch, A., Viramonte, J., Villarosa, G., and Salmuni, G.: Volcanic ash forecast during the June 2011 Cordón Caulle eruption, Nat. Hazards, 66, 389–412, https://doi.org/10.1007/s11069-012-0492-y, 2013.

Costa, A., Suzuki, Y., Cerminara, M., Devenish, B. J., Esposti Ongaro, T., Herzog, M., Van Eaton, A., Denby, L., Bursik, M., De' Michieli Vitturi, M., Engwell, S., Neri, A., Barsotti, S., Folch, A., Macedonio, G., Girault, F., Carazzo, G., Tait, S., Kaminski, É., Mastin, L., Woodhouse, M., Phillips, J., Hogg, A., Degruyter, W., and Bonadonna, C.: Overview of the Results of the Eruption Column Model Intercomparison Exercise, J. Volcanol. Geoth. Res., 326, 2–25, https://doi.org/10.1016/j.jvolgeores.2016.01.017, 2016.

Degruyter, W. and Bonadonna, C.: Improving on mass flow rate estimates of volcanic eruptions, Geophys. Res. Lett., 39, 1–6, https://doi.org/10.1029/2012GL052566, 2012.

Elissondo, M., Baumann, V., Bonadonna, C., Pistolesi, M., Cioni, R., Bertagnini, A., Biass, S., Herrero, J.-C., and Gonzalez, R.: Chronology and impact of the 2011 Cordón Caulle eruption, Chile, Nat. Hazards Earth Syst. Sci., 16, 675–704, https://doi.org/10.5194/nhess-16-675-2016, 2016.

Flentje, H., Claude, H., Elste, T., Gilge, S., Köhler, U., Plass-Dülmer, C., Steinbrecht, W., Thomas, W., Werner, A., and Fricke, W.: The Eyjafjallajökull eruption in April 2010 – detection of volcanic plume using in-situ measurements, ozone sondes and lidar-ceilometer profiles, Atmos. Chem. Phys., 10, 10085–10092, https://doi.org/10.5194/acp-10-10085-2010, 2010.

Folch, A.: A review of tephra transport and dispersal models: Evolution, current status, and future perspectives, J. Volcanol. Geoth. Res., 235–236, 96–115, https://doi.org/10.1016/j.jvolgeores.2012.05.020, 2012.

Folch, A., Costa, A., and Basart, S.: 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, 2012.

Folch, A., Costa, A., and Macedonio, G.: FPLUME-1.0: An integral volcanic plume model accounting for ash aggregation, Geosci. Model Dev., 9, 431–450, https://doi.org/10.5194/gmd-9-431-2016, 2016.

Galmarini, S., Bonnardot, F., Jones, A., Potempski, S., Robertson, L., and Martet, M.: Multi-model vs. EPS-based ensemble atmospheric dispersion simulations: A quantitative assessment on the ETEX-1 tracer experiment case, Atmos. Environ., 44, 3558–3567, https://doi.org/10.1016/j.atmosenv.2010.06.003, 2010.

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

Grell, G. and Baklanov, A.: Integrated modeling for forecasting weather and air quality: A call for fully coupled approaches, Atmos. Environ., 45, 6845–6851, https://doi.org/10.1016/j.atmosenv.2011.01.017, 2011.

Gudmundsson, M. T., Thordarson, T., Höskuldsson, Á., Larsen, G., Björnsson, H., Prata, F. J., Oddsson, B., Magnússon, E., Högnadóttir, T., Petersen, G. N., Hayward, C. L., Stevenson, J. A., and Jónsdóttir, I.: Ash generation and distribution from the April-May 2010 eruption of Eyjafjallajökull, Iceland, Sci. Rep., 2, 1–12, https://doi.org/10.1038/srep00572, 2012.

Institute of Earth Sciences: Eruption in Eyjafjallajökull 2010, available at: http://earthice.hi.is/eruption_eyjafjallajokull_2010 (last access: 21 December 2017), 2010.

Janjic, Z. and Gall, R.: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1 Dynamics, NCAR Technical Note, 2012.

Jay, J., Costa, F., Pritchard, M., Lara, L., Singer, B., and Herrin, J.: Erratum to "Locating magma reservoirs using InSAR and petrology before and during the 2011–2012 Cordón Caulle silicic eruption", Earth Planet. Sc. Lett., 395, 254–266, https://doi.org/10.1016/j.epsl.2014.07.021, 2014.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Edn., John Wiley & Sons, Ltd, Chichester, UK, 2012.

Jorba, O., Dabdub, D., Blaszczak-Boxe, C., Pérez, C., Janjic, Z., Baldasano, J. M., Spada, M., Badia, A., and Gonçalves, M.: Potential significance of photoexcited NO_2 on global air quality with the NMMB/BSC chemical transport model, J. Geophys. Res., 117, D13301, https://doi.org/10.1029/2012JD017730, 2012.

Marti, A., Folch, A., Costa, A., and Engwell, S.: Reconstructing the plinian and co-ignimbrite sources of large volcanic eruptions:

A novel approach for the Campanian Ignimbrite, Sci. Rep., 6, 21220, https://doi.org/10.1038/srep21220, 2016.

Marti, A., Folch, A., Jorba, O., and Janjic, Z.: Volcanic ash modeling with the online NMMB-MONARCH-ASH v1.0 model: model description, case simulation, and evaluation, Atmos. Chem. Phys., 17, 4005–4030, https://doi.org/10.5194/acp-17-4005-2017, 2017.

Mastin, L. G., Guffanti, M., Servranckx, R., Webley, P., Barsotti, S., Dean, K., Durant, A., Ewert, J. W., Neri, A., Rose, W., Schneider, D., Siebert, L., Stunder, B., Swanson, G., Tupper, A., Volentik, A., and Waythomas, C. F.: A multidisciplinary effort to assign realistic source parameters to models of volcanic ash-cloud transport and dispersion during eruptions, J. Volcanol. Geoth. Res., 186, 10–21, https://doi.org/10.1016/j.jvolgeores.2009.01.008, 2009.

Miller, T. P. and Casadevall, T. J.: Volcanic ash hazards to aviation, in: Encyclopedia of Volcanoes, edited by: Sigurdsson, H., Academic Press, Cambridge, 2000.

Mosca, S., Graziani, G., Klug, W., Bellasio, R., and Bianconi, R.: A statistical methodology for the evaluation of long-range dispersion models: an application to the ETEX exercise, Atmos. Environ., 32, 4307–4324, https://doi.org/10.1016/S1352-2310(98)00179-4, 1998.

Osores, M. S., Folch, A., Ruiz, J., and Collini, E.: Estimación de alturas de columna eruptiva a partir de imáges captadas por el sensor IMAGER del GOES-13, y su empleo para el pronóstico de dispersión y depóstio de cenizas volcánicas sobre Argentina, in XIX Congreso Geologico Argentino, 2014.

Oxford Economics: The economic impacts of air travel restrictions due to volcanic ash, available at: http://www.oxfordeconomics.com/my-oxford/projects/129051 (last access: 21 December 2017), 2010.

Prata, A. J. and Prata, A. T.: Eyjafjallajökull volcanic ash concentrations determined using Spin Enhanced Visible and Infrared Imager measurements, J. Geophys. Res.-Atmos., 117, 1–24, https://doi.org/10.1029/2011JD016800, 2012.

Prata, A. J. and Tupper, A.: Aviation hazards from volcanoes: the state of the science, Nat. Hazards, 51, 239–244, https://doi.org/10.1007/s11069-009-9415-y, 2009.

Stuefer, M., Freitas, S. R., Grell, G., Webley, P., Peckham, S., McKeen, S. A., and Egan, S. D.: Inclusion of ash and SO_2 emissions from volcanic eruptions in WRF-Chem: development and some applications, Geosci. Model Dev., 6, 457–468, https://doi.org/10.5194/gmd-6-457-2013, 2013.

Suzuki, T.: A theoretical model for dispersion of tephra, in: In Arc Volcanism: Physics and Tectonics, Terra Scientific Publishing Company, Tokyo, 93–113, 1983.

Webster, H. N., Thomson, D. J., Johnson, B. T., Heard, I. P. C., Turnbull, K., Marenco, F., Kristiansen, N. I., Dorsey, J., Minikin, A., Weinzierl, B., Schumann, U., Sparks, R. S. J., Loughlin, S. C., Hort, M. C., Leadbetter, S. J., Devenish, B. J., Manning, A. J., Witham, C. S., Haywood, J. M., and Golding, B. W.: Operational prediction of ash concentrations in the distal volcanic cloud from the 2010 Eyjafjallajökull eruption, J. Geophys. Res.-Atmos., 117, 1–17, https://doi.org/10.1029/2011JD016790, 2012.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Mon. Weather Rev., 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Wilkins, K. L., Watson, I. M., Kristiansen, N. I., Webster, H. N., Thomson, D. J., Dacre, H. F., and Prata, A. J.: Using data insertion with the NAME model to simulate the 8 May 2010 Eyjafjallajökull volcanic ash cloud, J. Geophys. Res.-Atmos., 121, 306–323, https://doi.org/10.1002/2015JD023895, 2016.

Wilson, T., Stewart, C., Bickerton, H., Baxter, P., Outes, V., Villarosa, G., and Rovere, E.: Impacts of the June 2011 Puyehue-Cordón Caulle volcanic complex eruption on urban infrastructure, agriculture and public health, 2013.

WMO: Commission for aeronautical meteorology – VAAC Inputs and Outputs Dispersion modelling Workshop, 2012.

Woodhouse, M., Hogg, A. J., Phillips, J. C., and Sparks, R. S. J.: Interaction between volcanic plumes and wind during the 2010 Eyjafjallajökull eruption, Iceland, J. Geophys. Res.-Sol. Ea., 118, 92–109, https://doi.org/10.1029/2012JB009592, 2013.