atmosphere

MDPI

Article

Machine Learning Emulation of Spatial Deposition from a Multi-Physics Ensemble of Weather and Atmospheric Transport Models

Nipun Gunawardena [1,*], Giuliana Pallotta [1], Matthew Simpson [2] and Donald D. Lucas [1,*]

[1] Lawrence Livermore National Laboratory, Livermore, CA 94550, USA; pallottagold1@llnl.gov
[2] Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA; mdsimpson@ucsd.edu
* Correspondence: gunawardena1@llnl.gov (N.G.); lucas26@llnl.gov (D.D.L.)

Abstract: In the event of an accidental or intentional hazardous material release in the atmosphere, researchers often run physics-based atmospheric transport and dispersion models to predict the extent and variation of the contaminant spread. These predictions are imperfect due to propagated uncertainty from atmospheric model physics (or parameterizations) and weather data initial conditions. Ensembles of simulations can be used to estimate uncertainty, but running large ensembles is often very time consuming and resource intensive, even using large supercomputers. In this paper, we present a machine-learning-based method which can be used to quickly emulate spatial deposition patterns from a multi-physics ensemble of dispersion simulations. We use a hybrid linear and logistic regression method that can predict deposition in more than 100,000 grid cells with as few as fifty training examples. Logistic regression provides probabilistic predictions of the presence or absence of hazardous materials, while linear regression predicts the quantity of hazardous materials. The coefficients of the linear regressions also open avenues of exploration regarding interpretability—the presented model can be used to find which physics schemes are most important over different spatial areas. A single regression prediction is on the order of 10,000 times faster than running a weather and dispersion simulation. However, considering the number of weather and dispersion simulations needed to train the regressions, the speed-up achieved when considering the whole ensemble is about 24 times. Ultimately, this work will allow atmospheric researchers to produce potential contamination scenarios with uncertainty estimates faster than previously possible, aiding public servants and first responders.

Keywords: deposition; machine learning; hazardous release; WRF; FLEXPART; prediction

Citation: Gunawardena, N.; Pallotta, G.; Simpson, M.; Lucas, D.D. Machine Learning Emulation of Spatial Deposition from a Multi-Physics Ensemble of Weather and Atmospheric Transport Models. *Atmosphere* **2021**, *12*, 953. https://doi.org/10.3390/atmos12080953

Academic Editor: Patrick Armand

Received: 15 June 2021
Accepted: 21 July 2021
Published: 24 July 2021

1. Introduction

From localized air-pollution caused by fireworks [1], to seasonal changes in pollution caused by cars [2], to planetary-scale dust transport from earth's deserts [3], particulate and gaseous hazardous matter can be dispersed throughout the environment from numerous natural and anthropogenic processes. One event which is important to public health and national security is the release of hazardous materials from nuclear weapons explosions, nuclear reactor breaches (such as Chernobyl or Fukushima), chemical spills, industrial accidents, and other toxic releases. These types of incidents happen suddenly and without warning, creating a plume of toxic material in the earth's atmosphere or ocean which can threaten the well-being of living organisms and environments.

In such situations, it is crucial that politicians, policy makers, and first responders have adequate knowledge about how the toxic plume will disperse and deposit throughout the environment. This can be used to determine evacuation zones and how resources are deployed to minimize the loss of public health. For example, during the 2011 Fukushima Daiichi disaster, the United States Department of Energy, the United States Environmental

Protection Agency, and other United States national agencies worked together to determine the effect of the radioactive release on international aviation routes, global food supply, and other crucial aspects of society [4].

To predict how a toxic plume disperses and deposits throughout the environment, scientists typically run computer simulations. These dispersion simulations solve physical and chemical equations to produce evolving concentration and deposition fields, but many of the processes represented in the models are uncertain or not resolved at the scales of interest. These processes are represented by empirical or semi-empirical parameterizations, and no single set of parameterizations always performs best for every scenario. Picking and choosing different sets of parameterizations provides an estimate of uncertainty and is a necessary component of the prediction process. In addition, many detailed transport and dispersion models that are currently in use are very computationally expensive, sometimes requiring several hours to complete a single simulation. Since time is of the essence during emergencies, these long run-times can be detrimental to first-responder efforts.

Therefore, in the event of a toxic environmental release, the scientists making predictions with limited computing resources often face a dilemma: using a detailed model, they can make a small number of predictions quickly but have poor knowledge of the uncertainty of those predictions, or they can make a large number of predictions slowly but have better knowledge of the uncertainty of the predictions.

A machine learning or statistical method that emulates a transport and dispersion model provides the opportunity to minimize this uncertainty versus time dilemma. To do this, the scientists would vary the inputs to the traditional weather/dispersion model to create a small number of predictions. They would then train a statistical model to produce dispersion predictions given the original input values. Finally, the statistical model could be used to create predictions for the set of inputs that were not originally run with the traditional dispersion model. That is to say, the statistical model is an emulator of the original dispersion model.

In this paper, we introduce a statistical method that rapidly predicts spatial deposition of radioactive materials over a wide area. The deposition predictions we are emulating were originally produced using material releases in the FLEXible PARTicle dispersion model (FLEXPART) [5] and meteorological fields generated from the Weather Research and Forecasting (WRF) [6] model. We created two FLEXPART-WRF ensembles for training and testing—one simulating a continuous surface release from a hypothetical industrial accident and one simulating an instantaneous elevated cloud from a hypothetical nuclear detonation. Each ensemble contained 1196 members with different weather conditions. (Each ensemble initially contained 1200 runs, but four runs did not complete due to numerical stability issues). To create the ensembles, WRF physics parameterizations were varied (i.e., a multi-physics ensemble) and used as inputs for our statistical model. Multi-physics WRF ensembles are often used to estimate weather model uncertainty, and our statistical method is able to capture this uncertainty very efficiently without having to run the full ensemble. We use a hybrid statistical model consisting of a two-dimensional grid of linear and logistic regression models for predicting spatial deposition.

The paper is organized as follows: Section 2 reviews the literature and the tools used. Section 3 describes the details of the dataset. Section 4 describes the statistical algorithm that is used as the emulator, and Section 5 presents the performance of the algorithm. Finally, Sections 6 and 7 discuss future work and summarize the current work, respectively.

2. Background

2.1. FLEXPART-WRF

There are many different methods that can be used to predict how an airborne contaminant disperses throughout the atmosphere, ranging from simple box models and Gaussian plumes to more sophisticated Lagrangian and Eulerian transport models [7]. Gaussian plume models are the simplest and the fastest to run but are often limited to very specific, idealized scenarios. Lagrangian and Eulerian models are slower but contain

representations of physical and chemical processes typically needed to simulate real-world releases. One key distinction between Gaussian plume models and Lagrangian/Eulerian models is that Gaussian plume models do not incorporate spatially and temporally varying meteorological fields.

We use the FLEXPART Lagrangian dispersion model for calculating the dispersion of airborne contaminants and estimate the effects of weather uncertainty in the dispersion calculations. To transport materials through the atmosphere and deposit them on the surface, FLEXPART requires spatially and temporally varying wind and precipitation data, which can come from archived meteorological forecast/analysis/re-analysis fields or weather models. For the work presented here, we use a specific version of FLEXPART designed to work directly with WRF output [5] (FLEXPART-WRF version 3.3). Although FLEXPART also has several internal physics options, we did not vary these for this work. A detailed description of our FLEXPART setup is provided in Section 3.

Several researchers have investigated the uncertainty of atmospheric dispersion models without incorporating machine learning. Leadbetter et al. [8] classify dispersion model error into three categories: source term uncertainty, meteorological uncertainty (which we study here), and intrinsic dispersion model uncertainty. They proceed to rank the uncertainties and find that wind direction and wind speed are important. Korsakissok et al. [9] ran multiple dispersion ensembles, some hypothetical and some realistic (e.g., the Fukushima Release) and analyzed the uncertainty. Finally, Sørensen et al. [10] simulated a nuclear power plant atmospheric release (similar to our surface release scenario) and presented a methodology to quantitatively estimate the variability of the ensemble. All studies cited the need for ensemble dispersion modeling. We focus specifically on uncertainty due to meteorological modeling.

To calculate winds and estimate weather uncertainty, we use the Weather Research and Forecasting model (WRF), a tool which is used to predict weather phenomena on scales of hundreds of meters to thousands of kilometers. A detailed description of WRF is found in Skamarock et al. [6]. WRF contains several physics options known as parameterizations for simulating processes such as cumulus convection, boundary layer turbulence, and land surface interactions. In our application, we estimate weather uncertainty by using a multi-physics approach that varies these parameterizations and uses the output to drive FLEXPART. A detailed description of the WRF setup is in Section 3. We specifically use WRF code version 3.9 with the advanced research dynamical numerical core.

Several other researchers have investigated WRF multi-physics ensembles. For example, researchers produced WRF multi-physics ensembles to investigate precipitation [11], heatwaves [12,13], and climate [14,15]. In prior work, we have investigated WRF multi-physics uncertainty to investigate the release from a nuclear power plant [16]. The important thing to note is that many of these ensembles have sizes of a few dozen members to a few hundred members. Our ensemble, having 1200 members, is feasible but significantly larger than average for a WRF multi-physics ensemble.

Machine learning and statistical methods have frequently been used to emulate complicated atmospheric models. Much of the prior emulation work focused on the potential speed-up offered, with applications to uncertainty studies, though some discussed ways to improve sub-grid scale parameterizations. Jensen et al. [17] and Lucas et al. [16] used machine learning algorithms to accelerate probabilistic inverse modeling studies of atmospheric sources. Watson [18] demonstrated the use of machine learning to improve long term climate statistics. Calbó et al. [19], Mayer et al. [20], and Beddows et al. [21] used polynomial chaos expansions and Gaussian processes to emulate air quality models. Wang et al. [22] used a neural network to emulate the planetary boundary layer parameterization of WRF. Krasnopolsky et al. [23] and Pal et al. [24] demonstrated the use of machine learning to emulate radiation parameterizations for global atmospheric models. Lucas and Prinn [25], Kelp et al. [26], and Ivatt and Evans [27] used statistical and machine learning approaches to emulate atmospheric chemistry and transport models. To our

best knowledge, this paper describes the first time a machine learning method is used to emulate full FLEXPART-WRF spatial deposition maps.

2.2. Linear and Logistic Regression

The two main statistical methods we used were linear regression and logistic regression. Since these simple methods are fast, easy to train, and readily interpretable, we used them over other more complex methods that we also investigated. Since linear regression and logistic regression are basic statistical tools, we will only present a brief overview here. More information about both methods can be found in many statistics and machine learning textbooks, such as Murphy [28] or Gelman and Hill [29]. Linear regression is a type of regression used to fit data that have a linear relationship. It can be written as $y = \boldsymbol{\beta}^T \mathbf{x}$, where y is the scalar output of the regression, $\boldsymbol{\beta}$ is an n-dimensional coefficient vector, T indicates the transpose operation, and $\mathbf{x}$ is the n-dimensional predictor vector. The first or last element in $\mathbf{x}$ is typically just set to 1 and is there so the fitted hyperplane has an intercept instead of being forced to pass through the origin. The "linear" in linear regression only applies to the coefficient vector—the elements of the input vector can be transformed as desired. Finally, linear regression without regularization is trained in a non-iterative fashion by minimizing the squared residuals. This contrasts with many other machine learning algorithms which are trained in an iterative fashion.

Logistic regression is a simple classification method that can be used to classify binary variables. It can be written as $p = \frac{1}{1+e^{-\boldsymbol{\beta}^T \mathbf{x}}}$. Here, p is the probability that the target class has a value of 1, $\boldsymbol{\beta}$ is an n-dimensional coefficient vector, T indicates the transpose operation, and $\mathbf{x}$ is the n-dimensional predictor vector. The function $f = \frac{1}{1+e^{-x}}$ is also known as the logistic function or sigmoid function. As with linear regression, logistic regression can have an intercept term. Unlike linear regression, logistic regression must be trained iteratively, even if regularization is not used.

3. Dataset

We trained our statistical model on two sets of FLEXPART dispersion simulations. Both sets release the radioactive particulate material cesium-137 (Cs-137), which has a half-life of 30.17 years, is highly soluble, and is subject to removal from the atmosphere by rainout and wet and dry deposition. Both sets of FLEXPART simulations use 1196 different weather conditions generated by a WRF multi-physics ensemble, as described below. The first set contains the results for a hypothetical continuous release of Cs-137 from the surface of the earth at an industrial facility representing a large-scale radiological accident. This set of simulations is referred to as the "surface release" case or the "surface" case. The second set contains simulations of a hypothetical instantaneous release of Cs-137 in the form of a mushroom cloud similar to how contaminants are created from a nuclear detonation. This set of simulations is referred to here as the "elevated release" case or "elevated" case. Any mathematical notation from this point forward can be generalized to either case unless otherwise specified.

Within a case, each ensemble member k consists of an 1×16 input vector $\mathbf{x}_k$ and an $M \times N$ target deposition map $\mathbf{Y}_k$, where M and N are the number of grid boxes in latitudinal and longitudinal directions, respectively. (The dimensionality of the input vector $\mathbf{x}_k$ will be explained later in this section). The vector $\mathbf{x}_k$ contains the physics parameterizations used by WRF and is the input to our statistical model. The deposition map $\mathbf{Y}_k$ is the output of FLEXPART-WRF given $\mathbf{x}_k$ and is used as the target data for training our statistical model. The input vectors are identical between the surface release case and the elevated release case because they are based on the same WRF ensemble, i.e., ${\mathbf{x}_k}^{\text{Surface}} = {\mathbf{x}_k}^{\text{Elevated}}$.

The FLEXPART settings remain constant for every ensemble member within a given case. Consequently, they are not included as inputs to our statistical model. Each FLEXPART simulation was set to begin at 12:00Z on 24 April 2018 and end 48 h later. An adaptive timestep was used for the sampling rate of the output, but the nominal value was 180 s. Subgrid terrain effects and turbulence were included, and land-use data were taken from

WRF. Two million Lagrangian particles were released, and the total mass for the surface and elevated cases was 1 kg and 0.28 g, respectively. We used the default precipitation scavenging coefficients for Cs-137. Table 1 shows the Cs-137 particle size distributions and masses as a function of altitude for the elevated release case, as further described in Norment [30]. Further information about the release scenarios can be found in Lucas et al. [31] and Lucas et al. [32].

While the FLEXPART settings of each ensemble member remain constant within the case, the set of physics options in WRF is different for every ensemble member. We vary the following five categories of physics parameterizations within WRF: planetary boundary layer physics (PBL), land surface model (LSM), cumulus physics (CU), microphysics (MP), and radiation (RA). Any remaining parameterizations or options remain fixed. To run WRF, one parameterization must be chosen from each physics category. While each category has several different parameterization options available, yielding well over 100,000 possible combinations of parameterizations, we selected a subset of 1200 possibilities expected to simulate the weather, as determined by expert judgment. The ensemble members were roughly chosen to maximize diversity in physics parameterizations.

In a real-world scenario, these 1200 possibilities would be forecasts, i.e., plausible scenarios for the time evolution of the weather and plumes over a two-day period given initial weather conditions that are known at the beginning of the forecast. Therefore, we assume that each ensemble member is equally likely and do not attempt to "prune" the ensemble while it is running because it is a short-term forecast. The 1200-member ensemble therefore provides an estimate of weather model uncertainty in forecasting the deposition from the hypothetical Cs-137 release events. Because we used data from 2018, we were able to verify the meteorological forecasts. In work not presented here, we ran simulations using data assimilation to produce analysis-observational fields. The ensemble simulations provide a reasonable spread around the nudged fields [32], which gives us confidence that our machine learning model can perform in realistic scenarios. Furthermore, for our short-term forecasts of two days, the WRF parameterization uncertainty is expected to dominate the variability. Very short term forecasts (e.g., 1 h) would not have a lot of variability, while longer forecasts (e.g., 7 days) have errors dominated by initial conditions, and the machine learning task would be much more difficult.

Ultimately, we selected five parameterizations for PBL, four for LSM, five for CU, four for MP, and three for RA. The specific parameterizations are shown in Table 2. This results in $5 \times 4 \times 5 \times 4 \times 3 = 1200$ different combinations of the WRF parameterizations. However, 4 of the 1200 combinations caused numerical issues in WRF, which failed to run to completion, so there are only 1196 members in the final multi-physics weather dataset. The 1196 input vectors $\mathbf{x}_k$ are vertically concatenated to create a 1196×16 input matrix $\mathbf{X}$. The 1196 output matrices $\mathbf{Y}_k$ are concatenated in the third dimension to make the $M \times N \times 1196$ output matrix $\mathbf{Y}$. The ordering of the parameterization combinations in the ensemble is shown in Figure 1.

The individual physics parameterizations are nominal categorical variables represented as numbers in WRF. In other words, the parameterizations are not ordinal—PBL parameterization 2, which represents the MYJ scheme, is not greater than PBL parameterization 1, which represents the YSU scheme. To prevent our statistical model from treating a higher numbered parameterization differently than a lower numbered parameterization, we transformed the input WRF parameterization vector using one-hot encoding [28]. This turns the five categorical variables for the parameterizations into sixteen boolean variables, which is why $\mathbf{x}_k$ has shape 1×16. For example, the LSM parameterization has four options: LSM 1, LSM 2, LSM 3, and LSM 7. When one-hot encoding, LSM 1 is represented by the vector $[0, 0, 0]$, LSM 2 is represented by the vector $[1, 0, 0]$, LSM 3 is represented by the vector $[0, 1, 0]$, and LSM 7 is represented by the vector $[0, 0, 1]$. The vectors for each parametrization are concatenated together. (For example, the ensemble member run with PBL 2, LSM 1, CU 5, MP 4, and RA 4 has a one-hot encoded input vector $[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1]$).

The output matrix $\mathbf{Y}$ consists of 1196 simulations produced by FLEXPART-WRF . Each ensemble member $\mathbf{Y}_k$ is an $M \times N$ map of the surface deposition of Cs-137 from either the surface release or the elevated release. For the surface release case, each map contains a total of 160,000 grid cells, with 400 cells in the latitudinal direction and 400 cells in longitudinal direction using a spatial resolution of about 1.7 km per cell. For the elevated release case, each map contains 600 grid cells by 600 grid cells with a resolution of about 1.2 km. Both deposition domains range from $+32.3^\circ$ to $+38.5^\circ$ in the latitudinal direction and -77.3° to -84.6° in the longitudinal direction. The height of the surface release domain was 3000 m resolved using 11 vertical layers, and the height of the elevated release domain was 4500 m resolved using 14 layers. The latitude and longitude of the location of the surface release were $+35.4325^\circ$ and -80.9483°, respectively. The latitude and longitude of the location of the elevated release were $+35.2260^\circ$ and -80.8486°, respectively. This domain is centered on the southwest corner of the US state North Carolina and has many different land types, including the Appalachian Mountains and the Atlantic Ocean.

The surface deposition output of FLEXPART-WRF accounts for both wet and dry removal of Cs-137 from the atmosphere and is reported in units of Bq/m^2 using a specific activity for Cs-137 of 3.215 Bq/nanogram. We also filtered out data less than 0.01 Bq/m^2 in our analysis, as deposition values below this level are comparable to background levels from global fallout [33] and do not pose much risk to public health.

All FLEXPART-WRF runs were completed on Lawrence Livermore National Laboratory's Quartz supercomputer which has 36 compute cores, 128 GB of RAM per node, and 3018 nodes total. A single WRF run costs about 150 core-hours, and a single FLEXPART run costs about 20 core-hours. The total ensemble cost was about 180,000 core-hours. The speedup between the full ensemble and the machine learning training set cost depends on the training size, which is discussed in Section 5. For a training size of 50, the total cost would be 7500 core-hours, which is a speedup of 24 times (or a savings of 172,500 core-hours). Figures 2 and 3 show selected examples of ensemble members from the surface case and elevated case, respectively. The members were chosen to highlight the diversity of the ensemble. The examples in the figures used PBL, LSM, CU, MP, and RA parameterization combinations (1, 1, 2, 3, 4) for member 25, (2, 1, 1, 3, 4) for member 245, (2, 3, 5, 3, 4) for member 413, and (7, 7, 10, 3, 4) for member 1157.

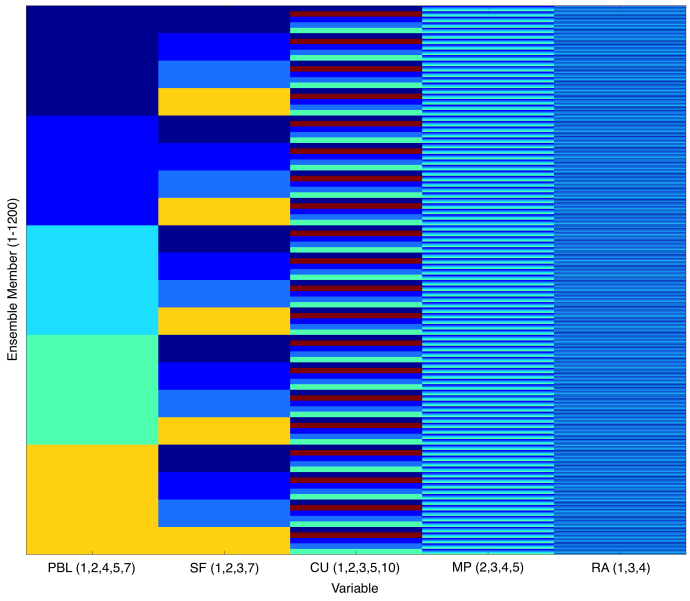

Figure 1. WRF parameterizations were varied as illustrated to create the multi-physics ensemble by iterating through the schemes in the order PBL, LSM, CU, MP, and RA.

Table 1. Profile for elevated release.

Altitude (m)	Mean Diameter (μm)	Geometric Standard Deviation	Cs-137 Mass (mg)
0	910	1.9	20.7
250	400	1.2	2.45
500	350	1.2	3.24
750	300	1.2	4.05
1000	270	1.2	5.17
1250	220	1.3	6.25
1500	170	1.4	9.67
1750	110	1.6	17.7
2000	52	2.1	51.5
2250	54	2.5	36.9
2500	48	2.4	36.6
2750	40	2.3	34.4
3000	32	2.0	26.1
3250	19	2.0	22.6

Table 2. WRF parameterizations used to create dataset, referred to here by their standard option number (between parentheses), name, and corresponding citation.

PBL	LSM	CU	MP	RA
(1) YSU [34]	(1) Thermal Diffusion [35]	(1) Kain-Fritsch [36]	(2) Lin (Purdue) [37]	(1) RRTM [38]
(2) MYJ [39]	(2) Noah [40]	(2) Betts-Miller-Janjic [39]	(3) WSM3 [41]	(3) CAM [42]
(4) QNSE [43]	(3) RUC [44]	(3) Grell-Devenyi [45]	(4) WSM5 [41]	(4) RRTMG [46]
(5) MYNN2 [47]	(7) Pleim-Xu [48]	(5) Grell-3 [49]	(5) Eta (Ferrier) [50]	
(7) ACM2 [51]		(10) CuP [52]		

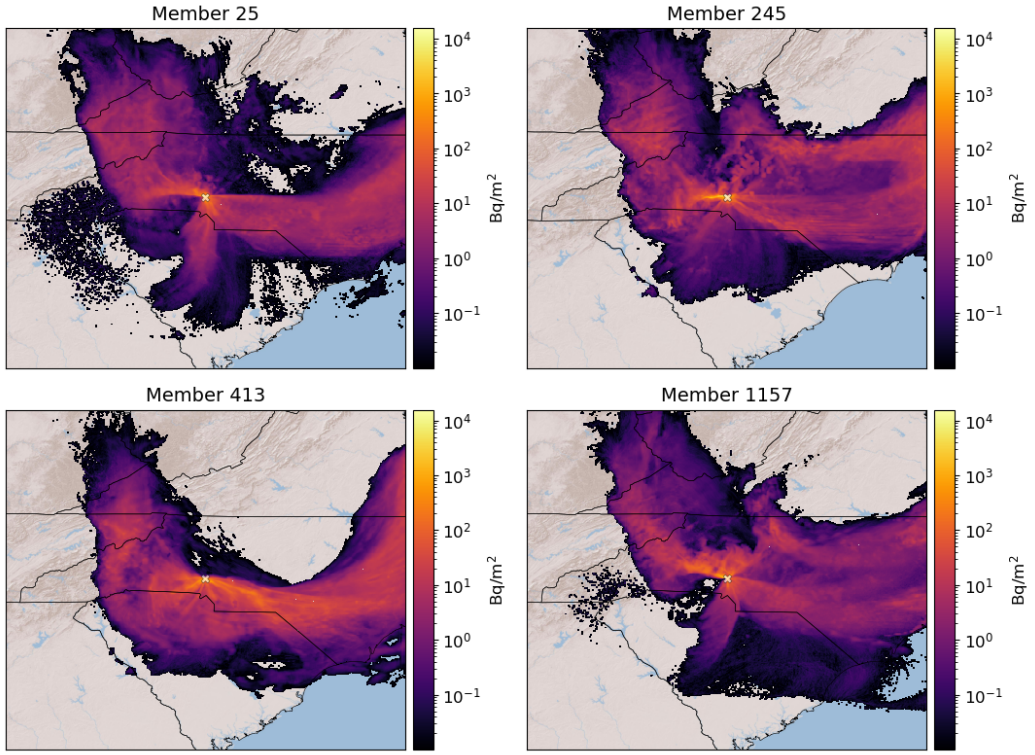

Figure 2. Examples of different deposition maps produced by FLEXPART-WRF for the surface release case. All values below 0.01 Bq/m^2 were removed. The WRF parameterizations used to create each subplot can be found in Section 3.

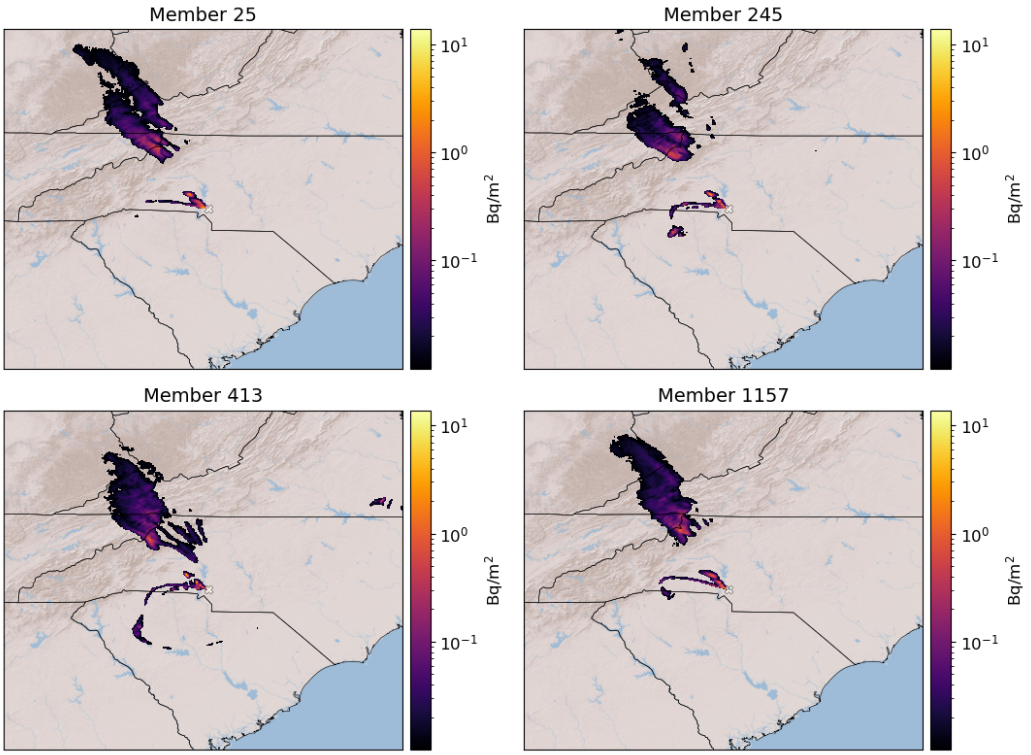

Figure 3. Examples of different deposition maps produced by FLEXPART-WRF for the elevated release case. All values below 0.01 Bq/m^2 were removed. The WRF parameterizations used to create each subplot can be found in Section 3.

4. Spatial Prediction Algorithm

The algorithm we use to emulate physics package changes in WRF is straightforward. A conceptual schematic can be seen in Figure 4. We start by creating an $M \times N$ grid $\hat{\mathbf{Y}}_k$ to represent the prediction of a given FLEXPART-WRF map $\mathbf{Y}_k$. Each grid cell $\hat{\mathbf{Y}}_{i,j,k}$ is the combined output of an independent linear regression and logistic regression model. The inputs to every linear and logistic regression model in the grid are the same: a 1×16 vector $\mathbf{x}_k$ of one-hot-encoded WRF physics categorical variables, as described in Section 3. For each grid cell, the associated logistic regression model determines the probability that the location will experience surface contamination from the hypothetical release event. If the probability at that location is greater than a pre-determined threshold value, the corresponding linear regression model determines the magnitude of the deposition. Mathematically, the value of a given grid cell $\hat{\mathbf{Y}}_{i,j,k}$ is given by Equations (1) and (2). The $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ terms represent the vector of regression coefficients for the logistic and linear regression models, respectively. The coefficients in Equation (2) are exponentiated because the linear regression is trained on the logarithm of the deposition. This linearizes the deposition values which allows the regression model to be fit; however, the logarithm is also useful for analyzing the data in general since the deposition values span many orders of magnitude.

$$\mathbf{P}_{i,j,k} = \frac{1}{1 + e^{-\boldsymbol{\alpha}_{i,j}^T \mathbf{x}_k}} \tag{1}$$

$$\hat{\mathbf{Y}}_{i,j,k} = \begin{cases} 0 & \text{if } \mathbf{P}_{i,j} \leq p_{\text{threshold}} \\ e^{\boldsymbol{\beta}_{i,j}^T \mathbf{x}_k} & \text{if } \mathbf{P}_{i,j} > p_{\text{threshold}} \end{cases} \tag{2}$$

The full $M \times N \times 1196$ dataset $\mathbf{Y}$ can be split into an $M \times N \times n$ training set $\mathbf{Y}_{\text{Train}}$ and an $M \times N \times (1196 - n)$ testing set $\mathbf{Y}_{\text{Test}}$. The linear regression models are trained on the logarithm of the deposition values, while the logistic regression models are trained on a binary indicator determining whether a grid cell has deposition or not.

We implemented our model in Python 3 using Numpy [53]. We used the linear regression and the logistic regression implementations from Scikit-Learn [54]. The logistic regression implementation in Scikit-Learn was run with the "liblinear" solver and L2 regularization with $\lambda = 1.0$. L2 regularization is necessary to obtain accurate results and ensure convergence. With 50 training examples, training regression models for every grid cell in the domain took approximately 1–1.5 min on a modern desktop computer. Making predictions for 1146 full maps took 5–6 min on the same computer, but that was achieved by re-implementing the Scikit-Learn "predict" functions using the Python just-in-time compiler Numba [55]. At approximately 315 ms per prediction on one core, the machine learning model offers an approximately two million times speedup for a single run. Some researchers have found similar speedups using ML on scientific codes [56]. Large scale experiments where the training and testing cycles had to occur thousands of times (e.g., determining training size convergence curves) were completed on Lawrence Livermore National Laboratory's Quartz Supercomputer and could take up to a few hours.

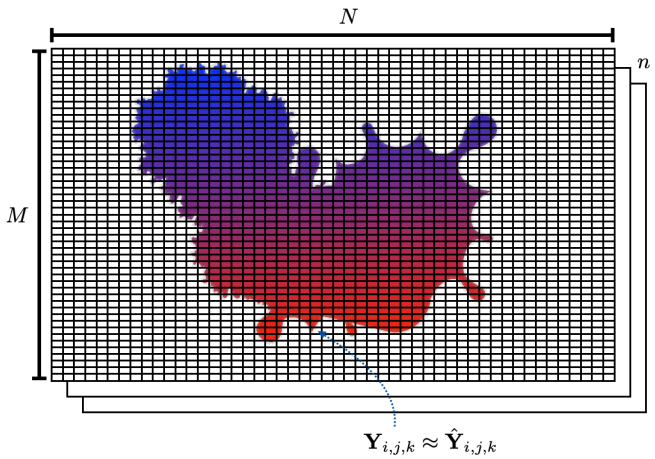

Figure 4. Conceptual diagram of the training set and model. Each deposition map produced by FLEXPART-WRF is a grid of size $M \times N$. There are n maps in the training set. A single grid cell of a single deposition map is represented by $\mathbf{Y}_{i,j,k}$ and can be approximated by our machine learning model output, $\hat{\mathbf{Y}}_{i,j,k}$. $\hat{\mathbf{Y}}_{i,j,k}$ is produced by the output of a single linear regression model and a single logistic regression model which is trained on the n data values for grid cell i, j. Further details of the model can be found in Section 4.

5. Results and Analysis

To test the effectiveness of our statistical model, we ran a suite of tests and derived performance statistics from the results. For these tests, we trained and evaluated our statistical model for eight different training sizes, with 100 runs with varying random seeds for each training size. The eight different training sizes we used were $n = 25$, $n = 50$, $n = 75$, $n = 100$, $n = 250$, $n = 500$, $n = 750$, and $n = 1000$ ensemble members. This corresponds to 2.09%, 4.18%, 6.27%, 8.36%, 20.90%, 41.81%, 62.71%, and 83.61% of our 1196-member ensemble dataset, respectively. Varying the random seed allowed each of the 100 runs for a given training size to have different members in the training set, which allowed us to see how much performance varied by training set member selection. The members of the test set for a given training size and random seed can be used in the training set for a different random seed. In other words, for a given training size and random seed, we had a training set and a testing set, but looking at all the random seeds for a given training size together was similar to k-fold cross validation. Since we used all 1196 members for this process, we did not have any truly held out test set that was not part of the 1196-member ensemble.

Figures that do not show training size variability (Figures 5–7) show the results from a 50-member training set with the same fixed random seed. The number 50 is somewhat

arbitrary but shows the minimum amount of training examples that produces accurate predictions. At 50 training examples, the predictions are qualitatively good, and one starts to see significant overlap between the training and testing performance metric distributions. Figures 8–10 all show results from the cross-validation tests.

The following subsections summarize the statistical and numerical performance of the algorithm. Some subsections present summary statistics, while some subsections present individual member predictions. In subsections where individual predictions are present, the training size is also presented.

5.1. Decision Threshold

Before showing summary statistics, it is important to understand how the output of our model is a probabilistic prediction. Figures 5 and 6 both have six subplots. The top left plot shows the true output by FLEXPART-WRF for a selected ensemble member. The top middle plot shows the probability map produced by the grid of logistic regression models. The color at each pixel represents the probability that the pixel has a non-zero deposition value. The areas of this subplot that are not colored are excluded from prediction because the corresponding grid cells in the training data contain no deposition. The remaining areas use the combination of logistic and linear regressions for making predictions.

The output of the logistic regression models is used in conjunction with a user-defined decision threshold value to produce deposition predictions. As determined from the training data, grid cells with probabilities greater than the threshold are predicted to have deposition, while those less than it are not. If conservative estimates are desired, a low threshold value can be used to include low probability, but still likely, areas of contamination in the prediction. The top-right and entire bottom row of Figures 5 and 6 show the predictions at different decision thresholds. The decision threshold can also be thought of as a probability cutoff value. The term "decision threshold" is synonymous with "decision boundary", which is referred to in the literature when classifying positive and negative outcomes [28].

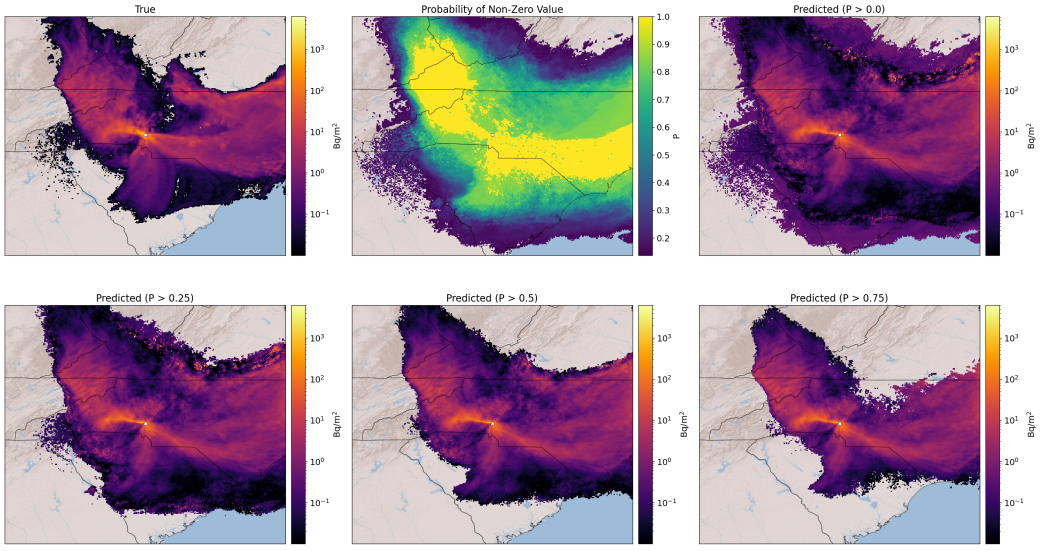

Figure 5. True FLEXPART-WRF output vs. predicted output at several decision threshold values for the surface release ensemble member 0 with $n = 50$. The WRF parameterization choices for this ensemble member were PBL 1, LSM 1, CU 1, MP 2, and RA 1. The top middle plot shows the original decision threshold map.

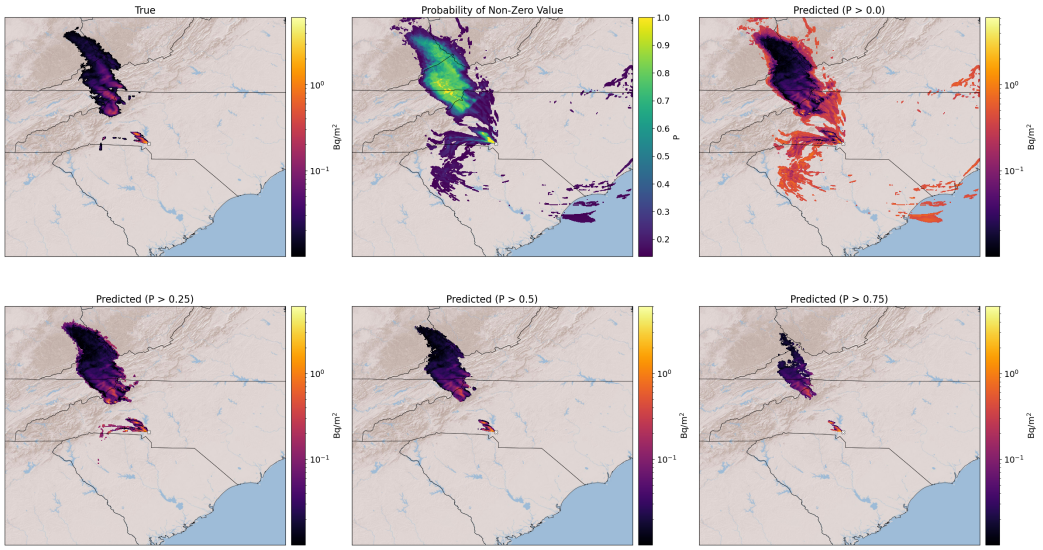

Figure 6. True FLEXPART-WRF output vs. predicted output at several decision threshold values for the elevated release ensemble member 0 with $n = 50$. The WRF parameterization choices for this ensemble member were PBL 1, LSM 1, CU 1, MP 2, and RA 1. The top middle plot shows the original decision threshold map.

Through a qualitative assessment, we determined that a decision threshold of 0.5 appears to be optimal. With values larger than 0.5, the plume shape starts becoming distorted and leaves important sections out. With values less than 0.5, noisy values at the edges of the plume are included, which are typically not accurate. These noisy values occur in grid cells where there are not many examples of deposition in the training data, and they are eliminated as more examples are included when the training size increases (see Section 5.2). These values can be seen in the bottom left subplot of Figure 5 on the northern edge of the plume. Anomalously large prediction values skew the performance statistics and are removed from the metrics if they exceed the maximum deposition value present in the training examples.

5.2. Training Size Variability

As with all statistical methods, the size of the training set affects the model performance. Figure 7 shows the plume prediction for ensemble member 296 as the training size increases. The members of the training set at a given size are also all included in the training set at the next largest size (i.e., the $n = 50$ training set is a proper subset of the $n = 75$ training set). The decision threshold is set to 0.5 for each training size. It is evident from the figure that as the training size increases, the deposition values and the plume boundary become less noisy. A quantitative assessment of how the predictions change with increasing training size is shown in Figures 8 and 9 for the surface case and elevated case, respectively.

These two figures show different statistical measures for predicting the members of training and testing sets as a function of training size. Because the selection of members is random and can affect the prediction performance, the experiment is repeated 100 times using different random seeds. Therefore, each "violin" in the plots displays the statistical variation stemming from member selection differences. For a given training size n, the orange training distributions are estimated from $n \times 100$ predictions, while the blue test distributions are derived from $(1196 - n) \times 100$ predictions.

The following error metrics are used to assess the predictive performance of the regression system. Two of the metrics target the logistic regressions (figure of merit in space and accuracy), three are for the linear regressions (fraction within a factor of 5, R, and fractional bias), and an aggregated metric (rank) is used to gauge the overall performance. Many other metrics are available to judge regression and classification

performance (e.g., mean squared error, F1), but we wanted to use metrics that were commonly used in the atmospheric science community [57,58].

- Figure of Merit in Space (FMS): A spatial error metric which is defined as the intersection of the area of the predicted and actual plumes divided by the union of area of the predicted and actual plumes [57]. Outside of atmospheric science this is also known as the Jaccard index. This metric depends only on the absence or presence of deposition, not the magnitude, and so directly assesses the logistic regressions. This metric varies between 0 and 1, and values of 0.8 and above are generally considered good for atmospheric models.
- Fraction within a Factor 5 (FAC5): The fraction of the predicted values within a factor of 5 of the actual values is an effective metric for assessing the linear regressions. Generalized, this is defined as FACX = Fraction of data that satisfy $\frac{1}{X} \leq \frac{\text{Predicted Value}}{\text{Actual Value}} \leq X$ [57]. We present the FAC5 value for the intersection of the predicted and actual plume locations. This metric can range from 0 to 1, with values above 0.85 generally being considered good values for atmospheric models.
- Pearson's R: Pearson's correlation coefficient R measures the linear relationship between the predicted and actual magnitudes of deposition in a log space. We present R calculated for the natural log of the intersection of the predicted and actual plume locations. This metric can range from -1 to 1, with values further away from 0 being good. (A Pearson's R value of -1 implies perfect anticorrelation, which is still useful).
- Accuracy: This is the standard classification accuracy as explained in Swets [59]. It is defined as the ratio of the sum of true positives and true negatives to all classifications considered (i.e., $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$). In a logistic regression case such as ours, a grid cell is classified as positive or negative by whether it has deposition above or below a certain threshold value. As described in Section 3, a deposition threshold of 0.01 Bq/m^2 was used. This metric can range from 0 to 1, with values closer to 1 being considered good.
- Fractional Bias (FB): Fractional bias is defined as FB $= \frac{\overline{C_O} - \overline{C_P}}{0.5(\overline{C_O} + \overline{C_P})}$, where $\overline{C_O}$ and $\overline{C_P}$ are the mean observed values and mean predicted values, respectively. It is a normalized indicator of model bias and essentially describes the difference in means in terms of the average mean. This metric ranges from -2 to +2 and has an ideal value of 0. In Figures 8 and 9, the fractional bias plot has a different shape from all the others. One reason for this is the fractional bias range is different, and the ideal value is in the center of the range. However, even if the absolute value of the fractional bias was taken, the shape would still be different. In this case, as the training set size increases, the fractional bias statistic converges to the inherent bias that exists between our statistical model and FLEXPART-WRF , just as the others do. However, in this case, it is shown that the training size that produces the least fractional bias is $n = 50$. This does not mean that $n = 50$ is the best sample size overall, as other metrics improve with increasing sample size. Like the other metrics, the fractional bias training and test curves converge with increasing training size, though they seem to converge much faster than the others.
- Rank: Described in Maurer et al. [58], the rank score is a metric that combines several statistics into a single number used to assess the overall performance of an atmospheric model. It is defined as Rank $= R^2 + (1 - \frac{|\text{FB}|}{2}) + \text{FAC5} + \text{Accuracy}$. R^2 is the coefficient of determination, and FB is the fractional bias. Each term in the equation ranges from 0 to 1, with 1 being best, which means the rank score ranges from 0 to 4, with 4 being best. The models studied in Maurer et al. [58] had a mean rank score of 2.06, which means our model looks very good in comparison. However, the models studied were time series models applied to individual stations, so they cannot be directly compared to our model.

In both the surface and elevated release cases, increasing the training size leads to, on average, an increase in performance on the test set and a decrease in performance on the training set. Nevertheless, as expected, the training set performance is better than the

testing set performance. There is no immediately distinguishable difference in performance between the surface case and the elevated case; on some metrics the surface case performs better and on others the elevated case performs better. However, the distribution of error metrics for the elevated case is often bimodal, whereas the surface case is more unimodal. This makes intuitive sense since the elevated case often has two separate deposition patterns with different shapes, while the surface case typically only has one large pattern.

Figures 7 and 8 highlight one of the most important conclusions from this work. Very few training samples are needed to make reasonable predictions. Even a prediction using 50 training samples, or $50/1196 = 4.18\%$ of the total dataset, is capable of accurately predicting deposition values in over 100,000 grid cells. Because there is significant overlap between the training and test distributions in Figure 8, these predictions are also robust to the 50 training samples selected from the full set.

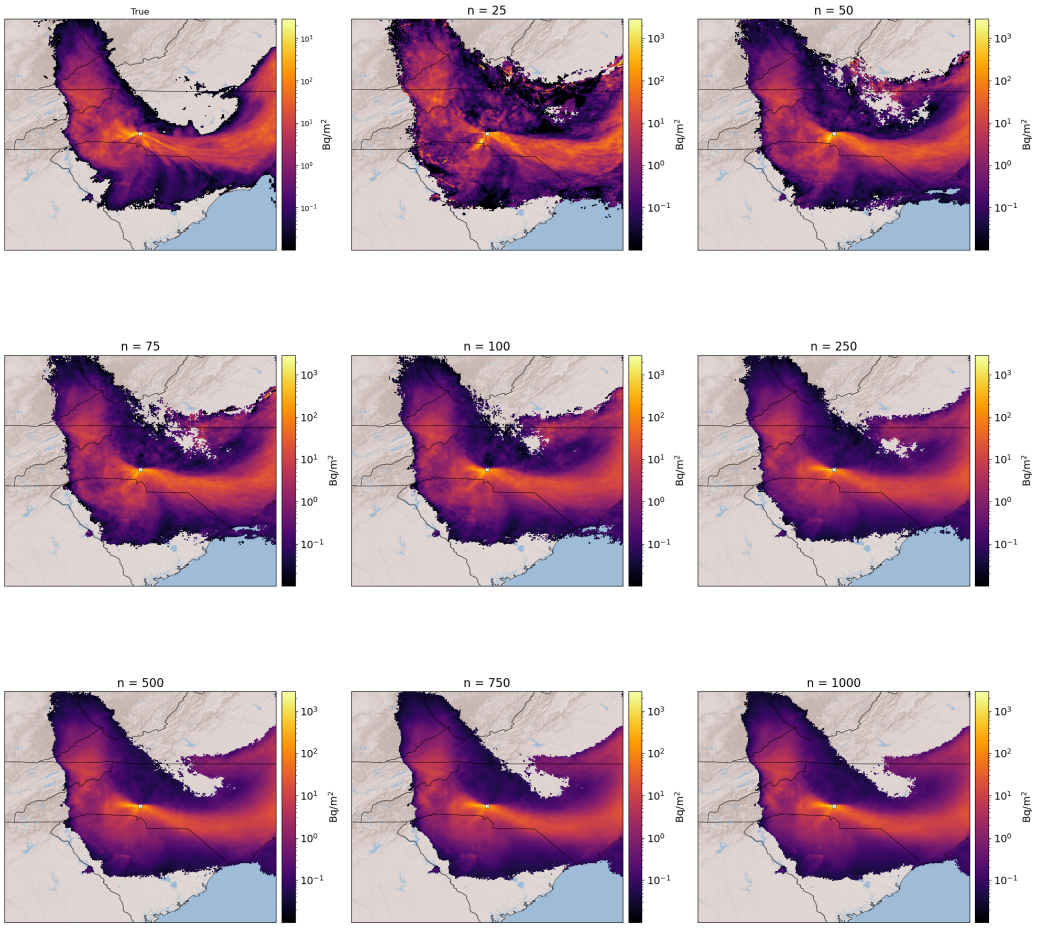

Figure 7. Spatial prediction for ensemble member 296 as the training set size increases. The true FLEXPART-WRF output is the top left subplot. The samples in the training set are randomly selected and the $p_{\text{threshold}}$ value is 0.5. Ensemble member 296 was in the test set for all training sizes. The WRF parameterization choices for this ensemble member were PBL 2, LSM 1, CU 5, MP 4, and RA 4.

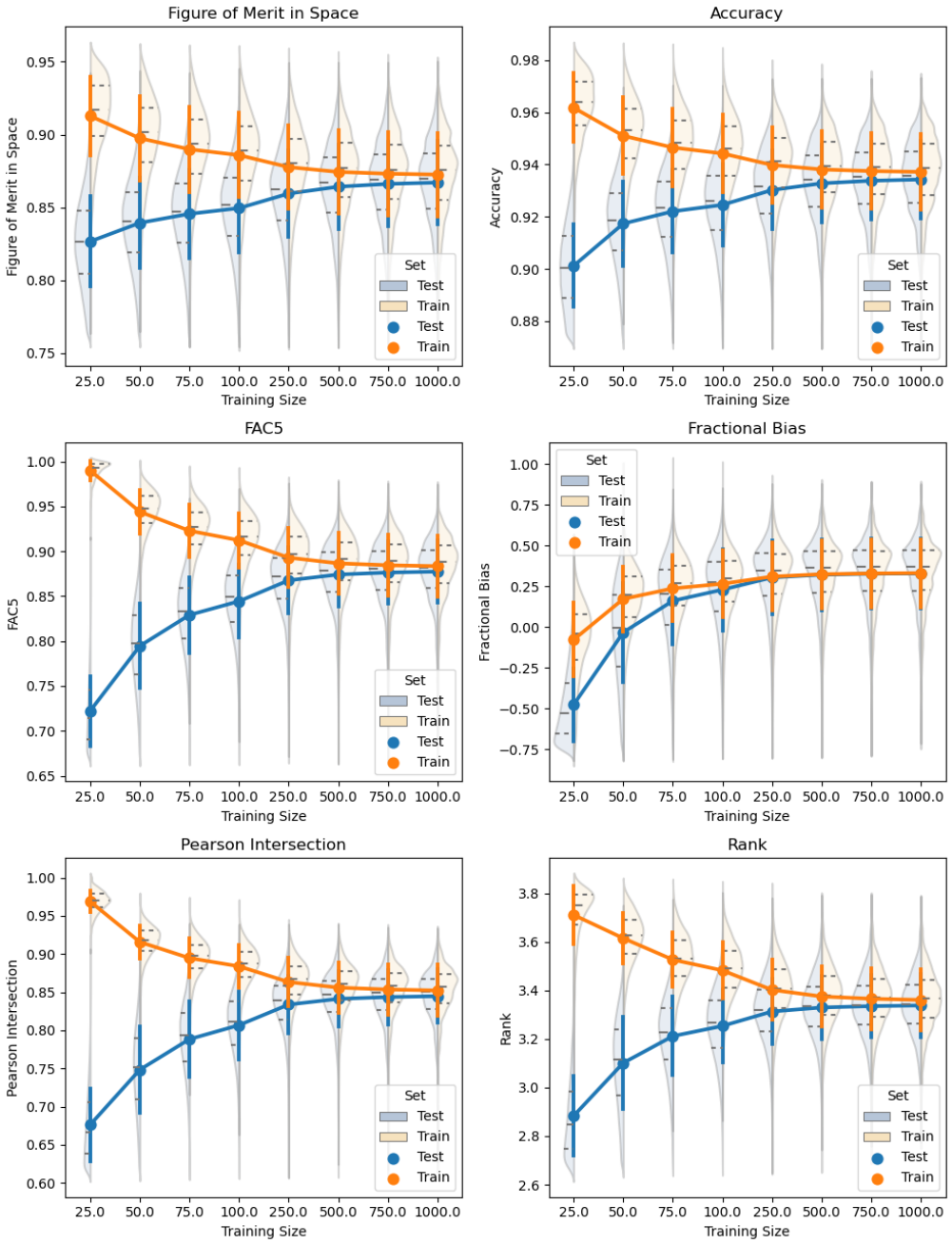

Figure 8. Spread of error metrics for all members for several different training sizes and 100 different random seeds for the surface release case. Training and test distributions are in blue and orange, respectively. A description of the metrics is provided in Section 5.2. Within the distributions, the dashed lines indicate the quartiles, the solid line is the mean, and the corresponding vertical bars are the standard deviations.

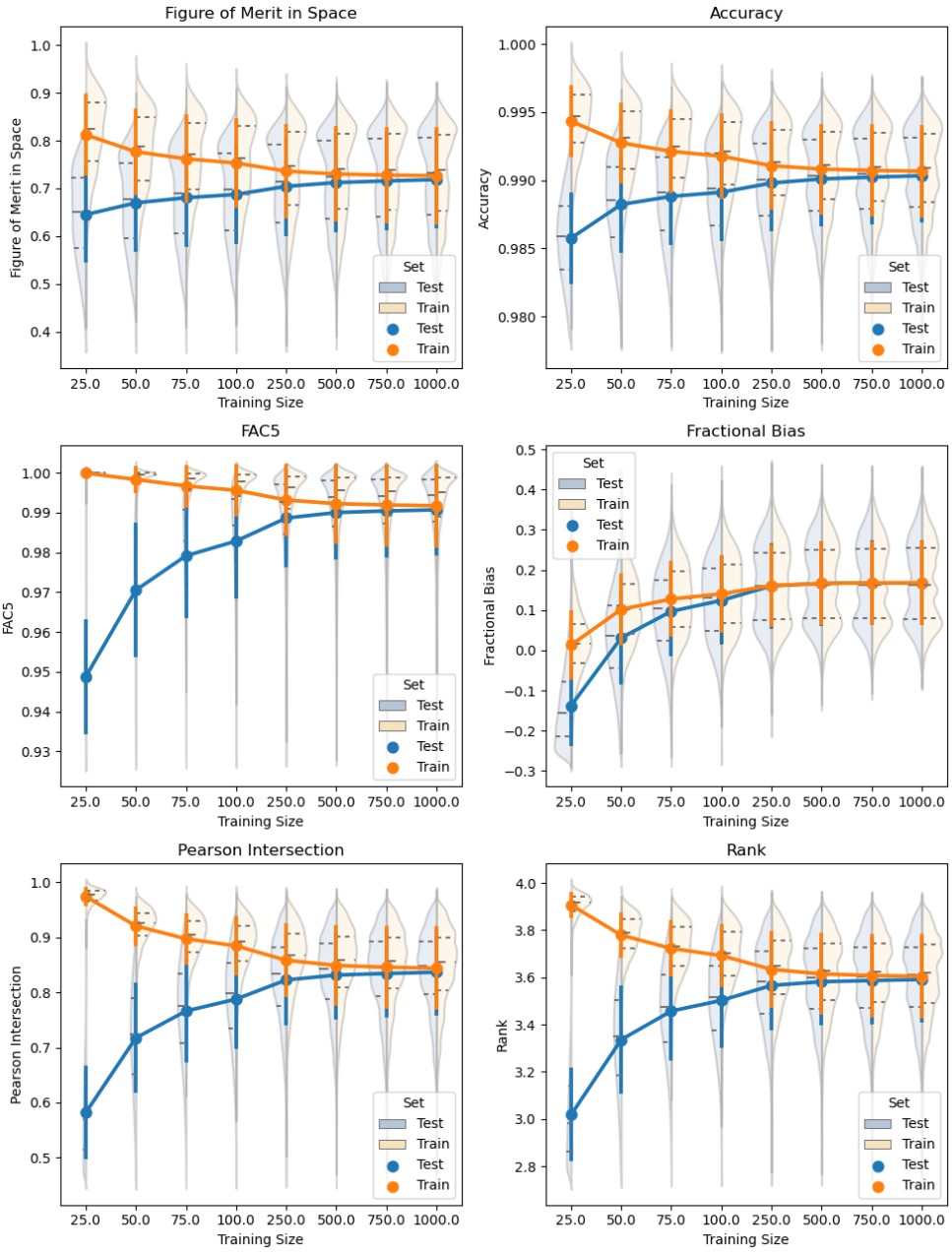

Figure 9. Same as Figure 8, except for the elevated release case.

5.3. Predictability of Individual Ensemble Members

The previous subsection described how training size affected the statistical model performance for the entire ensemble. In this section, we show how the predictions vary with training size for selected individual ensemble members. The purpose of this test is to show that some FLEXPART-WRF members are easier to predict than others, regardless of the amount of training members. Figure 10 shows the mean Pearson's R score by training size and member number for the surface release case for selected members of the test set. The members are selected by their decile average performance. We only show the members that are closest to the decile average performance because showing all 1196 members results in a visualization that is difficult to read.

For example, take the square marked by "60% (132)" on the x-axis and "250" on the y-axis. This square represents the mean Pearson's R score for member 132 calculated from every statistical model (out of 100) where member 132 was contained in the test set.

Member 132 is the member that is closest to the 60th percentile mean Pearson's R score averaged over *all* training sizes.

As already demonstrated, the general performance of the model increases as the training set size increases; however, the relative individual performance does not generally change. Part of this can be explained statistically. Our statistical model essentially fits a hyperplane in the WRF-parameter/deposition space. A hyperplane is one of the simplest possible models, and there is noise in the dataset. Some data points will be far away from the hyperplane, and increasing the training size does not move the hyperplane enough to successfully fit those points. This highlights the importance of the fact that physics-based modeling-machine learning is not able to capture all of the variation present in the dataset, even with very large training sizes. While we analyzed the WRF inputs associated with well and poorly performing members, we found no consistent pattern associated with poor predictions and WRF parameterizations. Hypothetically, if there was a relationship between WRF inputs and poorly performing members, the information could be used by WRF developers to improve accuracy for certain parameterizations. This figure also shows that low amounts of training data start producing accurate predictions. A similar analysis can be done for the elevated case but is not included here.

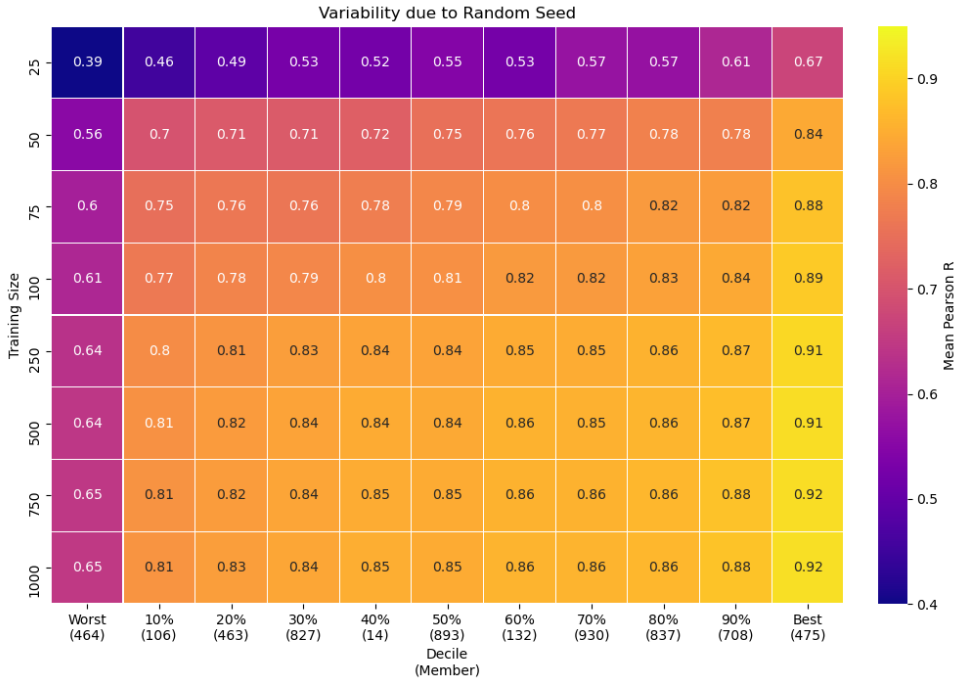

Figure 10. Mean Pearson R by training size and selected ensemble member. Some members are easier than others to predict regardless of the training size. Only instances where the ensemble member was included in the test set are used for calculations. The members were selected to be closest to the overall decile performance.

5.4. Ensemble Probability of Exceedance

One of the main goals of emulating Cs-137 spatial deposition is to account for the variability in the ensemble from weather uncertainty, so we use probability of exceedance plots to compare the variability of the predicted and true ensemble in Figure 11. The topmost and center subplots of Figure 11 show the percentage of members in the ensemble that have deposition values that exceed a threshold of 0.01 Bq/m^2 at every location. For example, if 598 ensemble members have deposition above 0.01 Bq/m^2 at grid cell (200, 200), the percentage for that cell is $598/1196 = 50\%$. Yellow colors indicate areas where many, if not all ensemble members report above-threshold deposition values. Dark purple colors indicate areas where very few ensemble members report above-threshold deposition values. Generally, the probability of exceedance drops as one moves further away from

the release location. The predictions are based on 50 training samples, and both ensembles used for this plot contain all 1196 members, meaning the training and testing predictions are included for the predicted percentages.

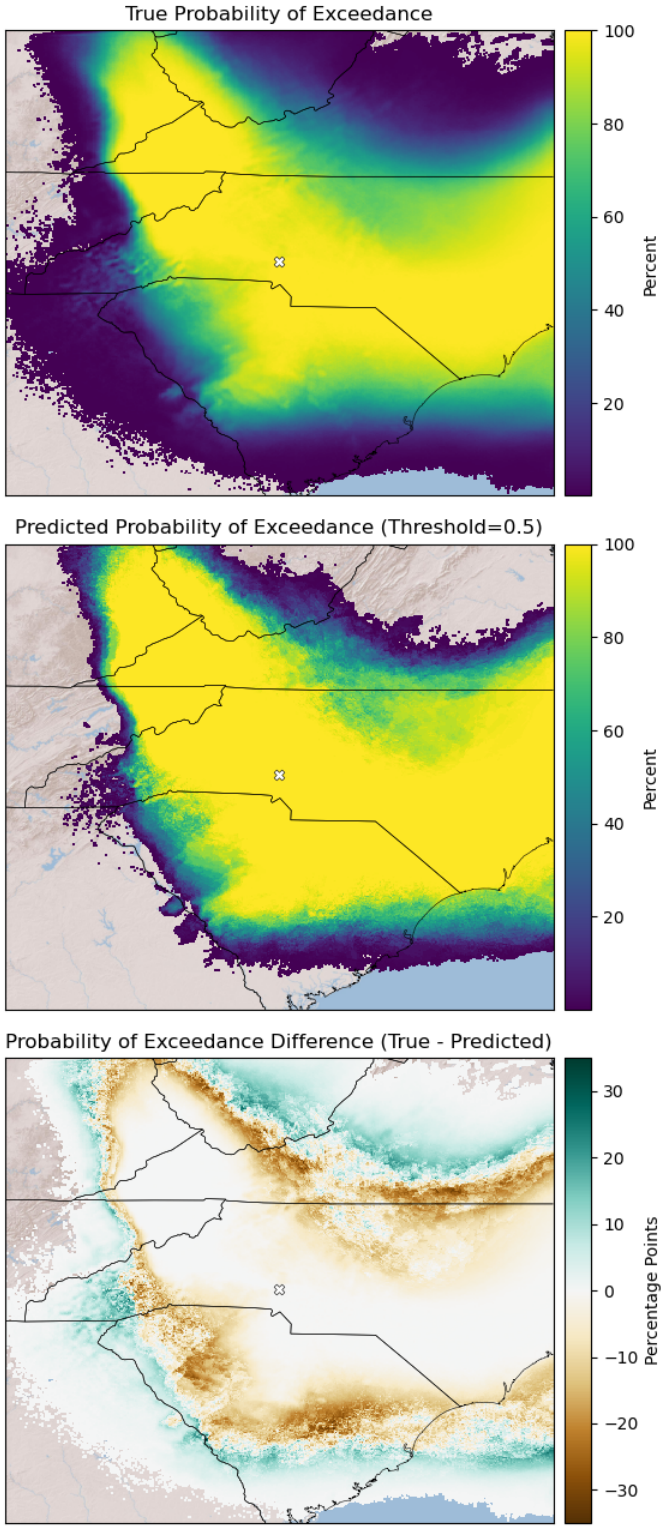

Figure 11. Percentage of members in the true (**top**) and predicted (**center**) ensembles that have deposition that exceeds 0.01 Bq/m^2 at each location. The bottom-most plot shows the difference between the two.

The topmost subplot shows the probability of exceedance of the true ensemble. As expected, the outside of the shape is made up of low percentage grid cells, as only outlier plumes make up those locations. The center subplot shows the probability of exceedance of the predicted ensemble. The predicted probability of exceedance takes up less area than the true ensemble because the outliers around the edge are challenging for the regressions to predict.

Despite the vast differences in computational resources needed to produce them, the probability of exceedance in the true and predicted ensembles appears similar. To highlight the differences, we created the bottom-most subplot of Figure 11, which shows the difference between the true ensemble percentages and the predicted ensemble percentages. Positive values, in teal, show areas where the population of members in the true ensemble is higher than the predicted ensemble. Negative values, in brown, show areas with higher predicted population than true population. Comparing the plot to Figure 12, one notices that the boundary between brown and teal happens approximately where the number of samples per pixel drops below 17, which is where the linear regression becomes underdefined. The conclusion we have drawn is that the regressions tend to overpredict values where there are sufficient samples (with some exceptions, such as in the center right of the plot) and underpredict where there are not sufficient samples.

5.5. Spatial Coefficient Analysis

One advantage our regression method holds over other machine learning models is the potential for interpretability. In this subsection we highlight one aspect of this interpretability. Our predictions are made using thousands of individual regression models, each of which has coefficients that transform the WRF parameterization input variables into a deposition value. In traditional regression approaches with non-categorical inputs, the units of all the input variables can be standardized so that the magnitude of a coefficient is related to the effect of its corresponding variable. That is, the larger the value of a coefficient, the more important the corresponding predictor is to the output. However, our WRF variables are one-hot-encoded as binary inputs, so determining their importance is not as straightforward as standard regression. Each of the regression models in our method has seventeen input terms—one for the intercept and sixteen binary encoded variables that represent five different WRF physics parameterizations. Out of these sixteen non-intercept coefficients, the first four represent the five PBL schemes, the next three represent the four LSM schemes, the next four represent the five CU schemes, the next three represent the four MP schemes, and the final two coefficients represent the three RA schemes. Taking the mean of the absolute value of a WRF physics package's coefficients gives an estimate of the importance of that variable. In other words, $\frac{1}{4}\sum_{i=1}^{4}|\beta_i|$ represents the importance of PBL, $\frac{1}{3}\sum_{i=5}^{7}|\beta_i|$ represents the importance of LSM, and so on.

Once the mean coefficient magnitudes are calculated, the argmax is used to find the WRF parameterization which is most important at a given grid cell. These results can be plotted to see which parameterizations are most important for a given area, as seen in Figure 12 for the surface release case. Figure 12 was created using models trained on 50 ensemble members and only includes grid cells that have greater than 17 samples. The intercept is not considered when determining importance. It is important to remember that with our process, the "most important variable" is not the same as "only important variable." Combinations of WRF parameterization changes can be important, resulting in the many coefficients that have a similar mean magnitude. In other words, the second most important WRF parameterization can still be very important because it has a mean coefficient magnitude slightly smaller than the most important WRF parameterization. Regardless, this analysis provides an interesting consequence of using regression models to interpret WRF physics.

Figure 12 shows that PBL variations tend to dominate other WRF parameterizations, as captured by the large areas in beige. This result is not surprising, as changing the PBL scheme in WRF is known to greatly influence atmospheric turbulence and mixing near the

surface. The variable importance map also shows other interesting features, including the red areas highlighting the relatively elevated importance of cumulus convection variations over coastal and mountainous areas where precipitation occurs during the release events. Similarly, magenta areas where microphysics is important occur near areas where cumulus convection is also important, which is consistent with the correlation of these physical processes in the model. The overall spatial complexity in Figure 12 illustrates one final critical point. No single WRF parameterization is most important everywhere, so multi-physics WRF ensembles that vary a range of physical parameterizations are needed to capture weather model uncertainty.

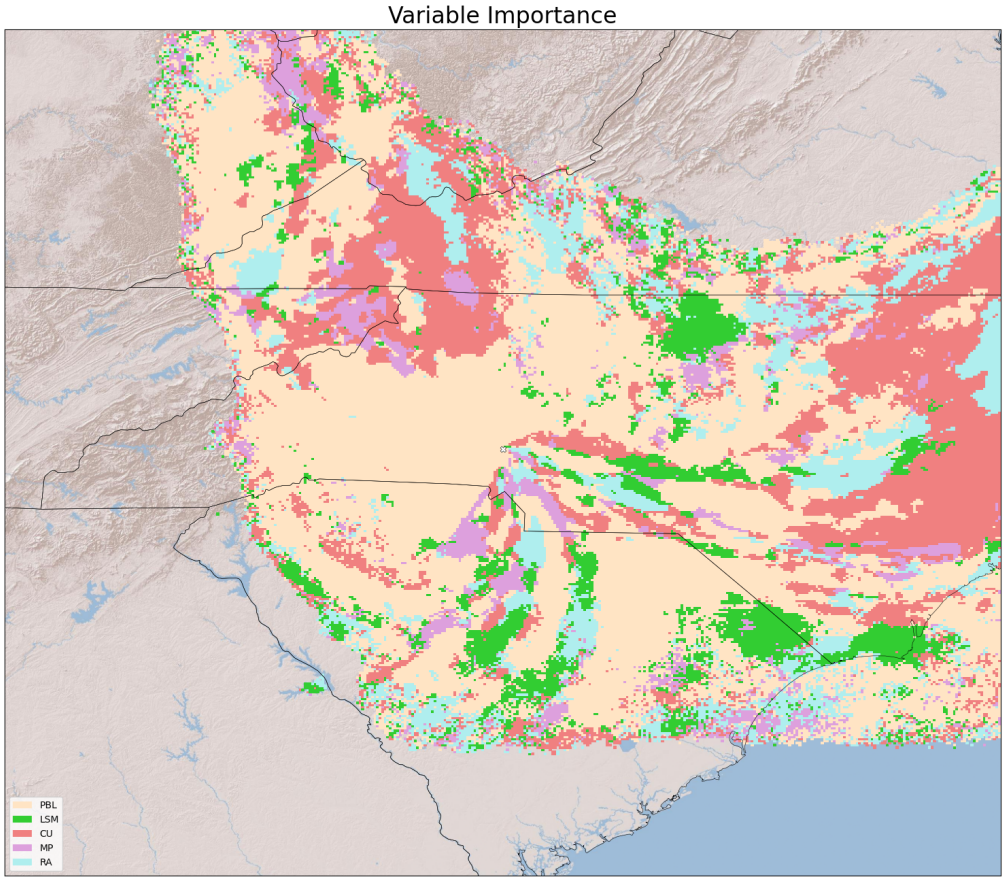

Figure 12. Primary WRF parameterization changes associated with Cs-137 deposition are color-coded and shown for every grid cell. Areas with 17 or fewer samples are excluded. The total training dataset included 50 ensemble members. In the legend, PBL stands for the planetary boundary layer physics parameterization (tan), LSM stands for land surface model (green), CU stands for cumulus physics (red), MP stands for microphysics (magenta), and RA stands for radiation (cyan).

6. Future Work

The regression prediction method we have described has some drawbacks and unknowns, which means there are several avenues for further exploration. The most significant drawback is that it does not exploit spatial correlations of nearby locations in the domain. Since each grid cell is treated as completely independent from the other grid cells, spatial correlations are not used to improve the overall prediction. This means that any predicted plume is limited to the envelope of the union of all of the training plumes, as our model cannot predict in areas that do not have any training data. However, this trait can be viewed as a positive feature of our algorithm; it will not poorly extrapolate in areas where there are no training data. To overcome this problem, spatial data can be incorporated into the overall model. Including spatial correlation information in our model may lead to a more parsimonious model or one that produces improved predictions. Including spatial

correlations can also potentially be done using dimensional reduction techniques such as PCA or autoencoders. For example, the model we describe could be used to produce an initial map, and then an alternate model based off radial basis functions, multitask learning, or even linear regression can be used to refine it.

Another drawback is the subjective nature of picking a decision threshold $p_{\text{threshold}}$ in the logistic regression component. We used a value of 0.5 for all the calculations presented here, which is a reasonable value to use, but that is the result of qualitative analysis. Implementing an optimization routine to determine the best $p_{\text{threshold}}$ to use would increase the objectivity and may improve the performance of our model. The tuned threshold could also be applied at a grid-cell level, which may increase performance in the boundary regions.

As mentioned in Section 5.1, we remove outlier deposition values which are predicted to be larger than any deposition value present in the training set. This is a simple way to remove outliers and is easily implemented operationally. However, it is a naive outlier removal method. A more complex outlier removal method may be beneficial to help differentiate false extreme values from true extreme values, the latter of which can pose a large risk to public health.

When we create training sets for our method we sample randomly from the entire population of predictions. By using methods from adaptive sampling, it may be possible to dynamically produce a training set that is more representative of the population than a random sample, leading to higher performance for the trained model with fewer expensive computer simulations. In an emergency situation, this would be very useful.

The individual models that predict hazardous deposition in each grid cell do not necessarily have to be linear or logistic regression models. They can be produced by other regression and classification models such as random forests or artificial neural networks. The biggest hurdle in implementing more complex grid cell-level models is the training time. During our testing on a desktop computer, the training time for a single grid cell took between 1 and 10 ms, and training a full spatial map was on the order of minutes. Changing to a more complicated model could potentially increase training time by an order of magnitude.

Finally, this regression method should be tested with more FLEXPART-WRF simulations. It should be tested with different hazardous releases in different locations from FLEXPART-WRF , but it could also be tested on completely different physical models. More terms could also be added to the regression model to account for larger initial condition errors present in longer forecast simulations. There is nothing about our method that is inherently specific to FLEXPART-WRF , and we think this method could work for simulations that are unrelated to deposition.

7. Conclusions

In this paper, we presented a statistical method that can be used to quickly emulate complex, spatially varying radiological deposition patterns produced by the meteorological and dispersion tools WRF and FLEXPART. FLEXPART-WRF is slow to run, and a single simulation from it may have significant uncertainty due to model imperfections. To estimate uncertainty, researchers can run FLEXPART-WRF hundreds of times by varying representations of physical processes in the models, but that can take crucial hours. Instead of running FLEXPART-WRF hundreds of times, researchers can run it dozens of times, use the results to train our emulator, and then use the emulator to produce the remaining results.

Our emulator is represented by an $M \times N$ grid where the value at each grid cell is determined by the output of independent linear regression and logistic regression models. The logistic regression determines whether hazardous deposition is present at that location, and the linear regression determines the magnitude of the deposition. Since all the grid cells are independent from one another, our model can accurately predict subsets of locations.

We used two datasets for training, testing, and predicting. One was a simulated continuous surface contaminant release representing a large-scale industrial accident, and the other was a simulated instantaneous elevated contaminant release from a hypothetical nuclear explosion. For each of the two cases, there were 1196 different simulations, all representing variations in the WRF parameterizations. The WRF parameterizations were treated as categorical variables that were binary encoded and used as the inputs to the linear and logistic regression models used in our emulator.

We conducted several tests to evaluate the performance of our emulator. We found that the emulator performs well, even with only 50 samples out of the 1196-member population. While the deposition patterns have variance, they are not drastically different shapes, which is why 50 samples is sufficient to make reasonable predictions. This is promising since in an emergency situation, the amount of computationally expensive runs should be minimized. As with many machine learning models, the prediction performance on the test set increases with increasing training size. We also found that for each case there are some members that perform better than others, regardless of the training size.

In general, we think that the emulator that we have presented here is successful in predicting complex spatial patterns produced by FLEXPART-WRF with relatively few training samples. We think there are several areas that can be explored to improve our emulator, and we hope to complete some of them in the future.

Author Contributions: N.G. contributed to statistical model preparation and analysis, visualization, and draft writing and editing. G.P. contributed to data analysis and validation. M.S. contributed to conceptualization, methodology, data creation, funding acquisition, and validation. D.D.L. contributed to statistical model analysis, data creation, draft writing and editing, validation, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. It was funded by LDRD 17-ERD-045. Released under LLNL-JRNL-808577.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The surface release data presented in this study are openly available at ftp://gdo148.ucllnl.org/pub/spatial, accessed on 22 July 2021. The elevated release data are available upon request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Notation

The following is a table of the notation used in the document. All notation is generalizable to the surface release case and the elevated release case.

Symbol	Meaning
$\mathbf{x}$	General single instance of predictor vector
$\mathbf{x}_k$	Specific single instance of predictor vector
$\mathbf{X}$	Complete set of input vectors
$\mathbf{X}_{\text{Train}}$	Training set of input vectors
$\mathbf{X}_{\text{Test}}$	Testing set of input vectors
$\mathbf{Y}$	Complete set of target matrices
$\mathbf{Y}_k$	Specific single instance of target matrix
$\mathbf{Y}_{i,j,k}$	Specific location of single instance of target matrix
$\mathbf{Y}_{\text{Train}}$	Training set of target matrices
$\mathbf{Y}_{\text{Test}}$	Testing set of target matrices
$\hat{\mathbf{Y}}$	Complete set of estimation matrices
$\hat{\mathbf{Y}}_k$	Specific single instance of estimation matrix

$\hat{\mathbf{Y}}_{i,j,k}$	Specific location of single instance of estimation matrix
M	Number of rows in $\mathbf{Y}$ or $\hat{\mathbf{Y}}$
N	Number of columns in $\mathbf{Y}$ or $\hat{\mathbf{Y}}$
n	Size of training set
$\boldsymbol{\alpha}_{i,j}$	Coefficients for the logistic regression model at a location
$\boldsymbol{\beta}_{i,j}$	Coefficients for the linear regression model at a location
$\mathbf{P}_{i,j}$	Probability of non-zero deposition at a location
$p_{\text{threshold}}$	Decision threshold to be applied to prediction

References

1. Moreno, T.; Querol, X.; Alastuey, A.; Minguillón, M.C.; Pey, J.; Rodriguez, S.; Miró, J.V.; Felis, C.; Gibbons, W. Recreational atmospheric pollution episodes: Inhalable metalliferous particles from firework displays. *Atmos. Environ.* **2007**, *41*, 913–922. [CrossRef]
2. Styer, P.; McMillan, N.; Gao, F.; Davis, J.; Sacks, J. Effect of outdoor airborne particulate matter on daily death counts. *Environ. Health Perspect.* **1995**, *103*, 490–497. [CrossRef] [PubMed]
3. Griffin, D.W.; Kellogg, C.A. Dust storms and their impact on ocean and human health: Dust in Earth's atmosphere. *EcoHealth* **2004**, *1*, 284–295. [CrossRef]
4. Bader, J.A. Dealing with Multiple Disasters in Japan. In *Obama and China's Rise: An Insider's Account of America's Asia Strategy*; Brookings Institution Press: Washington, DC, USA, 2012; pp. 130–139.
5. Brioude, J.; Arnold, D.; Stohl, A.; Cassiani, M.; Morton, D.; Seibert, P.; Angevine, W.; Evan, S.; Dingwell, A.; Fast, J.D.; et al. The Lagrangian particle dispersion model FLEXPART-WRF version 3.1. *Geosci. Model Dev.* **2013**, *6*, 1889–1904. [CrossRef]
6. Skamarock, W.C.; Klemp, J.B.; Dudhia, J.; Gill, D.O.; Barker, D.M.; Wang, W.; Powers, J.G. *A Description of the Advanced Research WRF Version 3*; NCAR Technical note-475+ STR; University Corporation for Atmospheric Research: Boulder, CO, USA, 2008.
7. Hutchinson, M.; Oh, H.; Chen, W.H. A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors. *Inf. Fusion* **2017**, *36*, 130–148. [CrossRef]
8. Leadbetter, S.; Andronopoulos, S.; Bedwell, P.; Chevalier-Jabet, K.; Geertsema, G.; Gering, F.; Hamburger, T.; Jones, A.; Klein, H.; Korsakissok, I.; et al. Ranking uncertainties in atmospheric dispersion modelling following the accidental release of radioactive material. *Radioprotection* **2020**, *55*, S51–S55. [CrossRef]
9. Korsakissok, I.; Périllat, R.; Andronopoulos, S.; Bedwell, P.; Berge, E.; Charnock, T.; Geertsema, G.; Gering, F.; Hamburger, T.; Klein, H.; et al. Uncertainty propagation in atmospheric dispersion models for radiological emergencies in the pre- and early release phase: Summary of case studies. *Radioprotection* **2020**, *55*, S57–S68. [CrossRef]
10. Sørensen, J.H.; Bartnicki, J.; Buhr, A.; Feddersen, H.; Hoe, S.; Israelson, C.; Klein, H.; Lauritzen, B.; Lindgren, J.; Schönfeldt, F.; et al. Uncertainties in atmospheric dispersion modelling during nuclear accidents. *J. Environ. Radioact.* **2020**, *222*, 106356. [CrossRef]
11. Kirthiga, S.; Narasimhan, B.; Balaji, C. A multi-physics ensemble approach for short-term precipitation forecasts at convective permitting scales based on sensitivity experiments over southern parts of peninsular India. *J. Earth Syst. Sci.* **2021**, *130*, 1–29. [CrossRef]
12. Imran, H.M.; Kala, J.; Ng, A.; Muthukumaran, S. An evaluation of the performance of a WRF multi-physics ensemble for heatwave events over the city of Melbourne in southeast Australia. *Clim. Dyn.* **2018**, *50*, 2553–2586. [CrossRef]
13. Stegehuis, A.I.; Vautard, R.; Ciais, P.; Teuling, A.J.; Miralles, D.G.; Wild, M. An observation-constrained multi-physics WRF ensemble for simulating European mega heat waves. *Geosci. Model Dev.* **2015**, *8*, 2285–2298. [CrossRef]
14. Katragkou, E.; García-Díez, M.; Vautard, R.; Sobolowski, S.; Zanis, P.; Alexandri, G.; Cardoso, R.M.; Colette, A.; Fernandez, J.; Gobiet, A.; et al. Regional climate hindcast simulations within EURO-CORDEX: Evaluation of a WRF multi-physics ensemble. *Geosci. Model Dev.* **2015**, *8*, 603–618. [CrossRef]
15. Lavin-Gullon, A.; Fernandez, J.; Bastin, S.; Cardoso, R.M.; Fita, L.; Giannaros, T.M.; Goergen, K.; Gutiérrez, J.M.; Kartsios, S.; Katragkou, E.; et al. Internal variability versus multi-physics uncertainty in a regional climate model. *Int. J. Climatol.* **2021**, *41*, E656–E671. [CrossRef]
16. Lucas, D.D.; Simpson, M.; Cameron-Smith, P.; Baskett, R.L. Bayesian inverse modeling of the atmospheric transport and emissions of a controlled tracer release from a nuclear power plant. *Atmos. Chem. Phys.* **2017**, *17*, 13521–13543. [CrossRef]
17. Jensen, D.D.; Lucas, D.D.; Lundquist, K.A.; Glascoe, L.G. Sensitivity of a Bayesian source-term estimation model to spatiotemporal sensor resolution. *Atmos. Environ. X* **2019**, *3*, 100045. [CrossRef]
18. Watson, P.A. Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *J. Adv. Model. Earth Syst.* **2019**, *11*, 1402–1417. [CrossRef]
19. Calbó, J.; Pan, W.; Webster, M.; Prinn, R.G.; McRae, G.J. Parameterization of urban subgrid scale processes in global atmospheric chemistry models. *J. Geophys. Res. Atmos.* **1998**, *103*, 3437–3451. [CrossRef]
20. Mayer, M.; Wang, C.; Webster, M.; Prinn, R.G. Linking local air pollution to global chemistry and climate. *J. Geophys. Res. Atmos.* **2000**, *105*, 22869–22896. [CrossRef]
21. Beddows, A.V.; Kitwiroon, N.; Williams, M.L.; Beevers, S.D. Emulation and Sensitivity Analysis of the Community Multiscale Air Quality Model for a UK Ozone Pollution Episode. *Environ. Sci. Technol.* **2017**, *51*, 6229–6236. [CrossRef]

22. Wang, J.; Balaprakash, P.; Kotamarthi, R. Fast domain-aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model. *Geosci. Model Dev.* **2019**, *12*, 4261–4274. [CrossRef]
23. Krasnopolsky, V.M.; Fox-Rabinovitz, M.S.; Chalikov, D.V. New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Mon. Weather Rev.* **2005**, *133*, 1370–1383. [CrossRef]
24. Pal, A.; Mahajan, S.; Norman, M.R. Using deep neural networks as cost-effective surrogate models for super-parameterized E3SM radiative transfer. *Geophys. Res. Lett.* **2019**, *46*, 6069–6079. [CrossRef]
25. Lucas, D.; Prinn, R. Parametric sensitivity and uncertainty analysis of dimethylsulfide oxidation in the clear-sky remote marine boundary layer. *Atmos. Chem. Phys.* **2005**, *5*, 1505–1525. [CrossRef]
26. Kelp, M.M.; Tessum, C.W.; Marshall, J.D. Orders-of-magnitude speedup in atmospheric chemistry modeling through neural network-based emulation. *arXiv* **2018**, arXiv:1808.03874.
27. Ivatt, P.D.; Evans, M.J. Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees. *Atmos. Chem. Phys.* **2020**, *20*, 8063–8082. [CrossRef]
28. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
29. Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Cambridge University Press: Cambridge, UK, 2006.
30. Norment, H.G. *DELFIC: Department of Defense Fallout Prediction System. Volume I-Fundamentals*; Final Report 16 Jan–31 Dec 79; Atmospheric Science Associates: Bedford, MA, USA, 1979.
31. Lucas, D.D.; Pallotta, G.; Simpson, M.D. Using Machine Learning to Intelligently Select Members of Large Atmospheric Model Ensembles. In Proceedings of the AGU Fall Meeting Abstracts, Washington, DC, USA, 10–14 December 2018; Volume 2018, p. GC43J-1663.
32. Lucas, D.D.; Simpson, M.; Pallotta, G. Probabilistic Predictions and Uncertainty Estimation Using Adaptively Designed Ensembles for Radiological Plume Modeling. In Proceedings of the CTBT Science and Technology 2019 Conference, Vienna, Austria, 24–28 June 2019.
33. Aoyama, M. Long-range transport of radiocaesium derived from global fallout and the Fukushima accident in the Pacific Ocean since 1953 through 2017—Part I: Source term and surface transport. *J. Radioanal. Nucl. Chem.* **2018**, *318*, 1519–1542. [CrossRef]
34. Hong, S.Y.; Noh, Y.; Dudhia, J. A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Weather Rev.* **2006**, *134*, 2318–2341. [CrossRef]
35. Dudhia, J. A multi-layer soil temperature model for MM5. In Proceedings of the Sixth PSU/NCAR Mesoscale Model Users' Workshop, Boulder, CO, USA, 22–24 July 1996; pp. 22–24.
36. Kain, J.S. The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteorol.* **2004**, *43*, 170–181. [CrossRef]
37. Chen, S.H.; Sun, W.Y. A One-dimensional Time Dependent Cloud Model. *J. Meteorol. Soc. Jpn. Ser. II* **2002**, *80*, 99–118. [CrossRef]
38. Dudhia, J. Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.* **1989**, *46*, 3077–3107. [CrossRef]
39. Janjić, Z.I. The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Weather Rev.* **1994**, *122*, 927–945. [CrossRef]
40. Tewari, M.; Chen, F.; Wang, W.; Dudhia, J.; LeMone, M.; Mitchell, K.; Ek, M.; Gayno, G.; Wegiel, J.; Cuenca, R. Implementation and verification of the unified NOAH land surface model in the WRF model. In *20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction*; American Meteorological Society: Seattle, WA, USA, 2004; Volume 1115, pp. 2165–2170.
41. Hong, S.Y.; Dudhia, J.; Chen, S.H. A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Weather Rev.* **2004**, *132*, 103–120. [CrossRef]
42. Collins, W.D.; Rasch, P.J.; Boville, B.A.; Hack, J.J.; McCaa, J.R.; Williamson, D.L.; Kiehl, J.T.; Briegleb, B.; Bitz, C.; Lin, S.J.; et al. Description of the NCAR community atmosphere model (CAM 3.0). *NCAR Tech. Note NCAR/TN-464+ STR* **2004**, *226*, 1326–1334.
43. Sukoriansky, S.; Galperin, B.; Perov, V. Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. *Bound.-Layer Meteorol.* **2005**, *117*, 231–257. [CrossRef]
44. Benjamin, S.G.; Grell, G.A.; Brown, J.M.; Smirnova, T.G.; Bleck, R. Mesoscale weather prediction with the RUC hybrid isentropic–terrain-following coordinate model. *Mon. Weather Rev.* **2004**, *132*, 473–494. [CrossRef]
45. Grell, G.A.; Freitas, S.R. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.* **2014**, *14*, 5233–5250. [CrossRef]
46. Iacono, M.J.; Delamere, J.S.; Mlawer, E.J.; Shephard, M.W.; Clough, S.A.; Collins, W.D. Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res. Atmos.* **2008**, *113*. [CrossRef]
47. Nakanishi, M.; Niino, H. An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteorol.* **2006**, *119*, 397–407. [CrossRef]
48. Gilliam, R.C.; Pleim, J.E. Performance assessment of new land surface and planetary boundary layer physics in the WRF-ARW. *J. Appl. Meteorol. Climatol.* **2010**, *49*, 760–774. [CrossRef]
49. Grell, G.A.; Dévényi, D. A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.* **2002**, *29*, 38-1–38-4. [CrossRef]
50. Rogers, E.; Black, T.; Ferrier, B.; Lin, Y.; Parrish, D.; DiMego, G. National Oceanic and Atmospheric Administration Changes to the NCEP Meso Eta Analysis and Forecast System: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3DVAR analysis. *NWS Tech. Proced. Bull.* **2001**, *488*, 15.

51. Pleim, J.E. A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *J. Appl. Meteorol. Climatol.* **2007**, *46*, 1383–1395. [CrossRef]
52. Berg, L.K.; Gustafson Jr, W.I.; Kassianov, E.I.; Deng, L. Evaluation of a modified scheme for shallow convection: Implementation of CuP and case studies. *Mon. Weather Rev.* **2013**, *141*, 134–147. [CrossRef]
53. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv* **2019**, arXiv:1907.10121.
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Lam, S.K.; Pitrou, A.; Seibert, S. Numba: A LLVM-based Python JIT Compiler. In Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, Austin, TX, USA, 15 November 2015; ACM: New York, NY, USA, 2015; pp. 1–6. [CrossRef]
56. Kasim, M.; Watson-Parris, D.; Deaconu, L.; Oliver, S.; Hatfield, P.; Froula, D.; Gregori, G.; Jarvis, M.; Khatiwala, S.; Korenaga, J.; et al. Building high accuracy emulators for scientific simulations with deep neural architecture search. *arXiv* **2020**, arXiv:2001.08055.
57. Chang, J.C.; Hanna, S.R. Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **2004**, *87*, 167–196. [CrossRef]
58. Maurer, C.; Baré, J.; Kusmierczyk-Michulec, J.; Crawford, A.; Eslinger, P.W.; Seibert, P.; Orr, B.; Philipp, A.; Ross, O.; Generoso, S.; et al. International challenge to model the long-range transport of radioxenon released from medical isotope production to six Comprehensive Nuclear-Test-Ban Treaty monitoring stations. *J. Environ. Radioact.* **2018**, *192*, 667–686. [CrossRef]
59. Swets, J.A. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285–1293. [CrossRef] [PubMed]