Revision of “Bayesian inverse modeling of the atmospheric transport and emissions of a controlled tracer release from a nuclear power plant”

Donald D. Lucas, Matthew Simpson, Philip Cameron-Smith, Ronald L. Baskett
Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA.

This merged PDF document includes:

1. Copy of response 1 containing red lines that show the areas of the marked-up manuscript that address reviewer 1's comments.

2.Copy of response 2 containing red lines that show the areas of the marked-up manuscript that address reviewer 2's comments.

3. Marked-up version of the manuscript showing differences between the reviewed and revised versions.

General response to Referee #1

We highly appreciate the feedback from Referee #1. The referee recommended a few minor technical corrections and commented on three issues. The issues are related to 1) the *mse* and *corr* metrics used to optimize model and measurement differences, 2) the apparent detrimental effects of nudging and data assimilation on the simulations, and 3) the information on modeling errors obtained by plugging in the known source values (location, time, and amount). We address these issues in further detail below and will revise the manuscript for clarity and content based on the reviewer feedback. Again, we thank the referee for the informative comments.

Minor comments

1. *mse and corr metrics*
   The reviewer asks about the advantage of fitting the mean-squared-error and correlation metrics instead of the actual concentrations for the optimization. The advantage is mainly one of statistical convenience. We could, in principle, fit statistical models to the actual concentrations (e.g., using gradient boosting), and then use the statistical models to minimize concentration differences with observations. Fitting the concentrations is more challenging, however, because the statistical model becomes functional and the size of the problem is larger. Referring to Eqs. 7 and 8 in the manuscript, our vectors and covariance matrix are only two dimensional (*mse* and *corr*). The terms in these equations would have more dimensions if we directly modeled the concentrations (1,148 without applying a dimensional reduction technique like principle components analysis). Further, we would have to estimate spatial and temporal correlations in the covariance matrix. We have collaborated with statisticians to tackle this more challenging statistical problem (Francom et al., 2016). This effort has shown that the additional complexity of modeling the concentrations can pay off in the form of tighter parameter estimates.

2. *Nudging effects*
   Like the reviewer, we find it counterintuitive that increased levels of meteorological nudging appear to degrade the agreement with concentration observations. Given the relatively small likelihood differences between the nudging options though, this result is not highly robust and not one of our major conclusions. Nonetheless, we also surmise that tradeoffs between nudging, variability, and physics parameterizations in complex terrain may be playing a role. At higher levels

of nudging, there is less variability, and therefore less opportunity for parameterizations to compensate for terrain-related errors. At low levels of nudging, there is more variability and a greater possibility to find better matches with observations. In future work it would be interesting to further diagnose nudging in complex terrain.

3. *Model error from known source*

   The reviewer recommends adding results using the known source parameters to help quantify model error. Both reviewers made the same excellent suggestion, so we expanded Sect. 5.1 in the revised manuscript to accommodate these results. A new scatterplot in the section shows the mean squared error and correlation for the 162 WRF runs with the known source parameters. These variations are not as large as they are across the full 40,000 member ensemble, though they are still significant. We also show that the known source variations are explained mainly by differences in the reanalysis data sets and initialization times, which agrees with our previous variance analysis in Sect. 5.3. Other factors, including nudging levels, do not play a very large role.

   See Sect. 5.1 and Figs. 7 and 8 in marked-up PDF.

Specific comments

In addition to the above comments, we will also revise the manuscript by

- providing a better describing of Latin hypercube sampling,

  See lines 379–391 in marked-up PDF.

- and improving figure captions (e.g. Fig. 3).

  See captions in marked-up PDF.

General response to Anonymous Referee #2

Anonymous referee #2 provided valuable feedback and comments about our ensemble analysis and inversion. The referee had six major comments, four minor comments, and recommended citing other studies using boosted regression trees (e.g. Sayegh et al). We greatly appreciate the reviewer's thoughtful comments, especially the suggestions to include additional information and analysis about the WRF variability using known source parameters. We address each of these comments below and have revised the manuscript accordingly. The revised manuscript includes descriptions and analysis for three new figures and modifications to an existing figure and table. We believe that these revisions are sufficient to address the reviewer's comments for publication in ACP.

Replies to Major Comments

1. *Analysis using actual release parameters*
   The reviewer recommended additional analysis using the known source parameters to highlight the variability due to WRF alone. The submitted version did not include this analysis because we wanted to show that the source inversion algorithm works well without knowledge of the source. That is, we excluded known source simulations from the training data as a rigorous test of our method. However the reviewer's recommendation is important and valuable, so we have expanded Sect. 5.1 in the revised manuscript to describe the known source analysis. This section includes two new figures showing the variability due only to WRF when the source parameters are known. One of the new figures, which is similar to Fig. 7 in the submitted version, shows the ensemble spread caused by the differences among the 162 WRF runs in the time series of SF_6 at the four selected locations. The other new figure shows the mean squared error and correlation for the 162 WRF runs using the known source parameters. Fig. 9 in the submitted version has also been modified to show the WRF-only variations relative to the larger Latin hypercube ensemble variations. The results of the new analysis are consistent with the findings described in the original manuscript. The WRF-only variations are important, though not as dramatic as the full ensemble variations. Further, the categorical variables in WRF related to reanalysis data and initialization time explain most of the spread within the known source WRF-only ensemble.

   See Sect. 5.1 and Figs. 7, 8, and 12 in marked-up PDF.

2. *Table of actual and estimated parameters*
   The reviewer requested a table that compares the source parameters estimated by the inversion algorithm to the actual values used for both the synthetic and real tracer releases. Table 3 in the submitted manuscript already shows the comparison for the real tracer release. This table is small and cross-referenced toward the end of the conclusions section, so it is not surprising that the reviewer missed it. In the revised manuscript, we expanded the table by adding the comparison for the synthetic tracer release and cross-reference it earlier in the discussion (in Sect. 5.5 and 5.6).

   See Table 3, line 1009, and line 1056 in marked-up PDF.

3. *Prior sampling ranges & insensitivity to release height*
   The reviewer notes that some of the sampling ranges for the source parameters in the prior distribution, especially the release height, seem to be too narrow because the posterior histograms extend to the limits. This behavior is not necessarily undesirable because the source parameters can be estimated from likelihood maxima that occur within the limits. Moreover the widths of the posterior distributions provide information about the uncertainty in estimated source parameters. Wide posterior distributions can result from observations that are not constraining or a model that is insensitive to parameter variations.

   For the release height, we used a narrow range to represent near-surface releases. We are not surprised that the simulations are not sensitive to the narrow variations because the reanalysis data sets and model simulations use a relatively coarse vertical resolution. We therefore expected to see similarities between the posterior and prior distributions for the release height (i.e., both are “flat”), which was confirmed by our results. It is often useful to include non-sensitive parameters in this type of Bayesian analysis to verify that the algorithm works properly.

   As the reviewer suggests, it would be interesting to broaden the sampling range for the release height to also cover elevated releases. However, this would require running a large number of new ensemble simulations, which we are unable to perform at this time (note that narrowing the range would not require new simulations). We will revise the manuscript to better describe the relationship between the posterior widths and parameter sensitivities, with an emphasis on the release height. We will also mention our interest to add elevated releases in future work.

   See changes in Sect. 3.2 and lines 1020-1039 in marked-up PDF.

4. *WRF resolution*

The 300-meter resolution of the innermost WRF domain was selected to resolve some of the terrain features surrounding Diablo Canyon. For instance, a line of measurement sensors detected SF_6 in a small canyon stretching from location 413 to the northwest (see Fig. 3). The selected model resolution is a compromise between being able to simulate the transport of SF_6 through these features and having scale-dependent parameterizations that are still valid.

Not all of the sensor locations require 300-m resolution, however, so the reviewer's suggestion to incorporate WRF resolution into the analysis is very interesting and relevant. We could potentially speed up the inversion by initializing with coarser WRF resolutions and progressing to finer resolutions. We could also quantify the joint effects of resolution differences and parameter variations. Though we haven't conducted this analysis for WRF and FLEXPART, we have done something similar for a global climate model. In that work, we ran ensembles at different resolutions and included resolution as an input parameter in our machine learning models. We will include additional discussion about WRF resolution in the revised manuscript. We also welcome future collaborations with the reviewer and others on this interesting topic.

See lines 163-169 in marked-up PDF.

5. *Number of observations*
The Diablo Canyon field study was heavily instrumented because of the complexity of the surrounding terrain and due to close proximity to populated areas. We agree that this data would be valuable for optimizing the effects of measurement density and location. Although we have conducted optimal analysis of this type for other studies and regions (e.g., see Lucas et al., 2015), we did not perform a similar optimization here. Instead, we roughly assessed the impacts of reduced data density by selecting fractions of measurement points at random, for fractions between 5% and 90%. We then re-computed the R^2 coefficients of determination for the gradient boosting regression models, as described in Sect. 5.4. As fewer measurement points are used in the fits, the quality of the GB models degrade, which, in turn, limits the ability to perform an inversion. From this analysis, using only 20% of the measurement points still results in R^2 values of about 0.82 and 0.96 for the *mse* and *corr*, respectively, which should be sufficient for estimating the most influential source parameters. Because some measurement locations are more important than others, we should be able to reduce the fraction of measurements below 20% without adversely affecting the inversion, if we optimize the locations. We will include a summary of this discussion in the revised manuscript and consider

future observational optimization studies for Diablo Canyon.

See lines 1264-1283 in marked-up PDF.

6. *Optimized versus regular parameters*

   We agree with the reviewer that it is a good idea to include a new figure that compares good and bad model simulations to the measurements. We selected high and low likelihood cases from the 40,000 member Latin hypercube ensemble and compare their SF_6 concentrations to measurements. This figure clearly shows that the high likelihood case provides a much better match to the measurements than the low likelihood case. It also shows that simulations have a positive bias, which leads to the minimum *mse* of about 0.2 instead of 0. This figure will be included and described in Sect. 5.4. Along with the modified version of Fig. 9 described in item 1 above, these figures will show the improvements that can be gained by varying WRF parameters and source parameters.

   See changes in Sect. 5.4 and Fig. 11 in marked-up PDF. Figs. 8 and 12 and Sect. 5.1 also contain information about the improvements due to changes only in WRF.

Replies to Minor Comments

As suggested by the reviewer, the revised manuscript will provide

- more details in the abstract about the optimized WRF configuration and accuracy of the parameter estimates,

  See abstract in marked-up PDF.

- an improved caption for Fig. 1,

  See Fig. 1 in marked-up PDF.

- further discussion about feature scores and parameter sensitivity,

  See lines 450-470 in marked-up PDF.

- and additional references and information about gradient boosting applications in atmospheric science.

  See lines 427-433 in marked-up PDF.

Bayesian inverse modeling of the atmospheric transport and emissions of a controlled tracer release from a nuclear power plant

Donald D. Lucas[a,*], Matthew Simpson[a], Philip Cameron-Smith[a], Ronald L. Baskett[a]

[a]*Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA.*

Abstract

Probability distribution functions (PDFs) of model inputs that affect the transport and dispersion of a trace gas released from a coastal California nuclear power plant are quantified using ensemble simulations, machine learning algorithms, and Bayesian inversion. The PDFs are constrained by observations of tracer concentrations and account for uncertainty in meteorology, transport, diffusion, and emissions. Meteorological uncertainty is calculated using an ensemble of simulations of the Weather Research and Forecasting (WRF) model that samples five categories of model inputs (initialization time, boundary layer physics, land surface model, nudging options, and reanalysis data). The WRF output is used to drive tens of thousands of FLEXPART dispersion simulations that sample a uniform distribution of six emissions inputs. Machine learning algorithms are trained on the ensemble data, and used to quantify the sources of ensemble variability and to infer, via inverse modeling, the values of the 11 model inputs most consistent with tracer measurements. We find a substantial ensemble spread in tracer concentrations (factors of 10 to 10^3), most of which is due to changing emissions inputs (about 80%), though the cumulative effects of meteorological variations are not negligible. The performance of the inverse method is verified using synthetic observations generated from arbitrarily selected simulations. When applied to measurements from a controlled tracer release experiment, the ~~most likely inversion results are within about~~ inverse method satisfactorily determines the location, start time, duration and amount. In a 2 km × 2 km area of possible locations, the actual location is determined to within 200 meters~~of the known release location, 5 and~~. The start time is determined to within 5 minutes out of 2 hours, and the duration to within 50~~minutes of the release start and duration times, respectively, and 22% of the release amount~~ minutes out of 4 hours. Over a range of release amounts of 10 kg to 1000 kg, the estimated amount exceeds the actual amount of 146 kg by only 32 kg. The inversion also estimates probabilities of different ~~combinations of WRF inputs of matching~~ WRF configurations. To best match the tracer observations, the highest probability cases in WRF are associated with using a late initialization time and specific reanalysis data products.

Keywords: meteorological and dispersion uncertainty, nuclear power plant emissions, source estimation, Bayesian inversion, machine learning

1. Introduction

Although the probability of a nuclear power plant accident is low, the risks associated with accidental releases of radioactive materials from nuclear power plants are expected to remain elevated worldwide through the coming decades (Christoudias et al., 2014). In the unlikely event of an accident, government agencies and plant owners must take actions to protect people and the environment from exposure to radioactive contamination. Because the atmosphere can spread the contaminants beyond the boundaries of a power plant within ~~a matter of~~ minutes to hours, reliable and timely protective action recommendations based on numerical modeling of actual releases are essential.

A variety of atmospheric models have been developed for simulating the transport and dispersion of releases from nuclear power plants, starting from the accidents at Three Mile Island in 1979 and Chernobyl in 1986 (e.g., Wahlen et al., 1980; Albergel et al., 1988; Gudiksen et al., 1989). These models range from simple straight-line Gaussian plumes that are applicable at short ranges when the turbulence in the atmo-

[*] Corresponding author
Email address: ddlucas@llnl.gov, ddlucas@alum.mit.edu (Donald D. Lucas)

Preprint submitted to Atmospheric Chemistry & Physics *September 11, 2017*

sphere is stationary and homogeneous (Seinfeld and Pandis, 2006), to more sophisticated models based on Lagrangian particles and/or Eulerian transport when the atmospheric flow is unsteady and occurs in areas with complex terrain (e.g., Pöllänen et al., 1997; Lauritzen and Mikkelsen, 1999; Nasstrom et al., 2007; Suh et al., 2009; Brioude et al., 2013a).

Atmospheric models used for nuclear power plant applications also use emissions modules to estimate the release rates of radionuclides based on specific reactor conditions (Athey et al., 1999). While these modules are useful for providing approximate ranges of emissions and their associated consequences, the detailed reactor conditions during an accident may not be well known and can contribute significant uncertainty to transport and dispersion predictions. The amount of radionuclides released to the atmosphere during the Fukushima Daiichi accident in 2011, for example, still remains highly uncertain because electrical power was lost and the reactors were monitored only indirectly (e.g., Chino et al., 2011; Terada et al., 2012; Stohl et al., 2012; Katata et al., 2015).

Inverse modeling can provide a safe way to infer information about radioactive emissions from nuclear power plants, and can also help estimate uncertainty in the meteorological fields used to transport the radioactive materials. Emissions and winds are constrained in an inverse method by minimizing differences between dispersion model predictions and observations of materials transported and deposited downwind from the source location (e.g., Davoine and Bocquet, 2007; Zheng and Chen, 2011). Building upon our previous work using Bayesian inverse modeling to estimate regional-scale greenhouse gas emissions (Lucas et al., 2015) and meteorological uncertainty in an urban-scale dispersion experiment (Lucas et al., 2016), we developed an ensemble-based inverse modeling system for analyzing nuclear power plant dispersion events. A diagram of the system is presented in Fig. (1) and summarized below.

Starting from the left hand side of the diagram in Fig. (1), an ensemble of plausible meteorological fields is generated using the Weather Research & Forecasting (WRF) model (Skamarock et al., 2008; Skamarock and Klemp, 2007). Ensemble members in WRF are created using different reanalysis datasets, physics packages, and configuration options, which are represented as categorical random variables. Sect. 2.1 provides further details of the WRF ensemble setup and design. The output of the WRF ensemble is then used to drive an ensemble of FLEXPART dispersion plumes (Brioude et al., 2013b), which also considers variations in the location, timing, and magnitude of emissions using continuous random variables. Further details of the FLEXPART calculations are given in Sect. 2.2.

The WRF-FLEXPART ensemble provides a set of plume predictions that are compared with field measurements (right hand side of Fig. 1). The differences between the simulations and field data are minimized through Monte Carlo sampling loops that jointly vary the inputs to WRF and emissions in FLEXPART (red dashed lines in the diagram). Because a single iteration through the sampling loops is computationally expensive and millions of iterations may be needed, we use machine learning to accelerate the optimization. Moreover, by sampling the WRF and FLEXPART inputs with a probability distribution function, Bayesian analysis is used to estimate uncertainty in the model inputs and outputs. Details of the machine learning and inversions methods are provided in Sects. 3.1 and 3.2.

Measurements from a tracer release experiment conducted (Thuillier, 1992) in September 1986 at the Diablo Canyon nuclear power plant are used to test and verify the atmospheric models and inversion system in Fig. (1). Diablo Canyon is located along the rugged coast of California (see Fig. 2), so the study provides a critical test of simulating transport and dispersion in complex terrain. The study also provides an important verification test of the inversion algorithm, because the tracer emissions are assumed to be unknown and inferred in the inversion. Information about the Diablo Canyon study is given in Sect. 4.

To our knowledge, this work represents the first joint inversion capability applied to FLEXPART dispersion simulations that provides probability distribution functions of categorical inputs in WRF and continuous inputs in FLEXPART, and has been verified with tracer release data. This capability can be useful for other applications beyond releases from nuclear power plants, including compliance monitoring of the nuclear test ban treaty (Issartel and Baverel, 2003) and inverse modeling of emissions from large-scale industrial accidents and volcanic plumes (e.g., Heng et al., 2016).

2. Ensemble atmospheric modeling

2.1. Weather Research & Forecasting Model

The non-hydrostatic, fully compressible Weather Research and Forecasting (WRF) atmospheric model (Skamarock et al., 2008; Skamarock and Klemp, 2007) is used to generate meteorological fields for the atmospheric transport and diffusion simulations. Version 3.6.0 of the advance research WRF (ARW) core is used for the simulations presented in this paper. WRF was developed through collaboration among government, research, and academic organizations to facilitate the transfer of state-of-the-science atmospheric research findings to an operational modeling capability. The National Center for Atmospheric Research currently maintains the open source WRF model code, which is ~~publically~~ publicly available for user download. WRF is widely used by numerous groups for both atmospheric research and real-time operational commercial applications, such as renewable generation and utility grid demand forecasting.

2.1.1. WRF Domain

High resolution winds are needed to simulate the dispersion for the Diablo Canyon tracer release test problem (see Fig. 2 and Sect. 4). Through efficient numerical model nesting and the parallelization of source code for high performance computers, WRF can be used to simulate a large range in scales of motion from thousands of kilometers down to tens of meters (Lundquist et al., 2010).

Using this nesting capability, five WRF model domains are used to downscale and generate high-resolution meteorological fields over the Diablo Canyon region. The WRF domain configuration and geographic coverage are shown ~~on the left hand side of~~ in Fig. 3. A large portion of the Western United States is covered by the outermost model domain (labeled D1) at 24.3 km horizontal grid spacing. The large outer model domain was required for downscaling purposes given the coarse resolution of some of the reanalysis data sets used to initialize WRF for this study.

By downscaling, a horizontal grid spacing of 300 m is achieved in the innermost WRF model domain (D5 ~~, right hand side of~~ in Fig. 3). The fine-scale grid spacing of D5 is necessary to generate representative meteorological conditions since the Diablo Canyon nuclear power plant is located on the coast near narrow valleys and other topographic features that result in complex terrain induced flow. Some of the sensors shown in Fig. 3 are located in places requiring even finer resolution (e.g., the line of sensors that stretch to the northwest of site 413), but the physics parameterizations in WRF are not designed for such small scales. The 300-m grid spacing in D5 is therefore a trade-off between resolution and model physics. High-resolution terrain and land use fields were generated for D5 by downloading National Elevation Dataset and National Land Cover Database at 1-arc second (approx. 30 m) data from the United States Geological Survey Multi-Resolution Land Characteristics data server (Homer et al., 2015).

2.1.2. WRF Ensemble

Several features also make WRF ideal for creating an ensemble of plausible atmospheric conditions for uncertainty assessments. Ensemble modeling approaches have been shown to be effective at quantifying physically plausible states of the atmosphere in a probabilistic manner (e.g., Mullen and Baumhefner, 1994; Stensrud et al., 2000; Berner et al., 2011). The WRF modeling system contains numerous physics schemes that parameterize subgrid-scale processes, such as the surface energy exchange, cloud microphysics, and turbulent mixing in the planetary boundary layer. By running WRF with ~~multiple~~ combinations of physics options, an ensemble is generated that captures meteorological uncertainty due to subgrid-scale parameterization error. In addition, WRF can be initialized and run using a variety of ~~publically~~ publicly available reanalysis data sets to quantify meteorological uncertainty resulting from errors in initial / lateral boundary conditions. Meteorological data from gridded analysis data sets and observations can also be integrated into WRF simulations to improve model accuracy by using four dimensional data assimilation (FDDA) options for analysis (Stauffer and Seaman, 1994) and observational (Liu et al., 2005, 2009) nudging. The goal of the analysis FDDA option is to nudge large-scale motion towards an observed state using relaxation terms, while the observational nudging impacts the prediction of local-scale atmospheric phenomenon.

Table 1 summarizes the five major variables that were selected for the WRF ensemble. Reference "base" values for each variable are also listed in the table. The base values were selected using expert judgment and represent a prior WRF configuration expected to perform well. For the purposes of the Monte Carlo sampling and analysis (Fig. 1), the WRF variables are treated as categorical random variables. By taking all of the combinations among the five variables in the table, we constructed a WRF ensemble containing 162 members. The major categories in the table include model initialization time, source of input reanalysis data, FDDA nudging weighting factors, planetary boundary layer (PBL) physics, and land surface model (LSM) physics. The ensemble categories and their variations were selected based on previous WRF user experience and a literature review of sources of uncertainty that are likely to impact meteorological fields important to the specific tracer release experiment described in Sect. 4. Variations in PBL and LSM physics, for example, have been shown to affect near surface stability and wind fields (Lee et al., 2012), which can have a large impact on plume transport modeling. Other categories related to changes in microphysics and cumulus physics are not considered in this report because precipitation and cloud cover were not present during the specific tracer release experiment. Future studies can easily incorporate these factors, and others, by including additional categories for Monte Carlo sampling.

Several variations are included in the weather ensemble to account for uncertainty related to model initialization and meteorological reanalysis inputs. WRF simulations were started at either 15 hours or 9 hours before the beginning of the tracer release (i.e., at 00:00 UTC or 06:00 UTC on 4 September 1986) to investigate the sensitivity of model solutions to initialization start time and model spin up duration. All of the WRF simulations ended at 13 hours after the end of the tracer release (i.e., at 12:00 UTC on 5 September 1986). The three reanalysis variations included are the North American Regional Reanalysis (NARR) data (Mesinger et al., 2006), European Centre for Medium-Range Weather Forecasts (ECMWF) data (Hersbach et al., 2015), and Climate Forecast System Reanalysis (CFSR) fields (Saha et al., 2010). NARR reanalysis meteorological data are available every 3 hours on 30 vertical levels with a horizontal grid spacing of 32 km. Both ECMWF and CFSR reanalysis fields are available every 6 hours on 38 vertical levels. However, CFSR reanalysis fields have a horizontal grid spacing of roughly 60 km versus roughly 125 km for ECMWF data.

The FDDA weighting of meteorological data fields during the WRF ensemble simulations was varied to account for uncertainty associated with the assimilation of gridded reanalysis fields and irregularly spaced weather observations. Weather simulations were performed with WRF FDDA options for analysis and observational nudging options either turned off, using default weighting factors as suggested by WRF guidance, or a high option with the weighting factors one order of magnitude higher than the default values. Additionally, FDDA analysis nudging was used only on the two outer course-resolution model domains, while FDDA observational nudging was used only on the two innermost model domains (see WRF domains in Fig. 3). FDDA observation nudging included surface METAR measurements and multi-level data from the backup and primary meteorological towers at Diablo Canyon (see Sect. 4).

Three PBL models and three LSM schemes were used to construct the WRF ensemble to account for uncertainty associated with turbulent mixing and surface momentum, moisture, and thermodynamic fluxes. The PBL models included the Yonsei University (YSU) scheme (Hong et al., 2006), the Mellor-Yamada-Janjic (MYJ) scheme (Janjić, 1994), and the Mellor-Yamada-Nakanishi and Niino (MYNN) scheme (Nakanishi and Niino, 2006). Among the PBL models, the biggest difference is that the YSU scheme uses a countergradient flux (non-local) method to develop parabolic mixing profiles in the boundary layer, while the MYJ and MYNN ~~schmes~~ schemes use different numerical approaches to solve for local turbulent kinetic energy (TKE) based vertical mixing in the PBL and free atmosphere. The LSM physics options include Thermal Diffusion (Duhdia, 1996), NOAH (Ek et al., 2003), and RUC (Benjamin et al., 2004) models. Soil moisture and explicit vegetation canopy physics are not included in the Thermal Diffusion model, while the NOAH and RUC models parameterize vegetation canopy effects to differing degrees and both provide soil moisture gradients.

2.2. *FLEXPART*

The FLEXPART Lagrangian dispersion particle model (Stohl et al., 1998, 2005; Stohl and Thomson, 1999) was used to simulate the atmospheric transport and mixing of the tracer gas released from the Diablo Canyon nuclear power plant. Field experiments have been used to validate the performance of FLEXPART (Stohl et al., 1998; Forster et al., 2007). FLEXPART has also been used in a wide variety of dispersion applications, including the transport of air pollutants (An et al., 2014; Avey et al., 2007), of radiological releases from nuclear power plants and radioisotope production facilities (Andreev et al., 1998; Wotawa et al., 2010; Stohl et al., 2012), of volcanic plumes (Stohl et al., 2011), and of noble gases produced from nuclear weapons tests (Becker et al., 2010).

We used FLEXPART-WRF version 3.1 (Brioude et al., 2013b), which was developed to use meteorological data generated by the WRF model to drive atmospheric transport and diffusion processes. The FLEXPART-WRF code is open source and available for download (https://www.flexpart.eu/). Mean particle trajectories and tracer concentrations were calculated using the three dimensional wind components from WRF (u, v, and w) over the 50 km by 50 km domain shown in Fig. 3 (dashed rectangular area). The FLEXPART-WRF grid used 401 cells in each of the horizontal directions and 11 vertical levels from the surface to 3 km, with 6 levels contained in the lowest 500 m. Tracer concentrations were derived using one million particles released from a randomly selected point source location at random release times, as detailed in the next section. Wind fluctuations (σ_v and σ_w) were calculated using parameterizations (Hanna, 1982; Ryall and Maryon, 1998) and WRF micrometeorological output variables (friction velocity, surface sensible heat flux, planetary boundary layer height, and Monin Obukhov length scale). Lagrangian particles were evolved using a sampling rate and synchronization interval of 20 seconds, and the simulations utilized a subgrid terrain parameterization. FLEXPART-WRF can also simulate wet and dry deposition removal processes (Wesely and Hicks, 1977; McMahon and Denison, 1979; Slinn, 1982; Hertel et al., 1995), though these processes were not needed for the passive gas tracer release at Diablo Canyon.

2.2.1. FLEXPART Ensemble

In addition to the wind field variations generated by the WRF ensemble, the inverse modeling system in Fig. 1 also applies a Monte Carlo sampling loop to emissions variables in FLEXPART. The goal of this part of the inversion is to determine the location, timing, and magnitude of the tracer release emissions by minimizing the differences between FLEXPART predictions and field measurements.

The location, timing, and magnitude of the Diablo Canyon release are inferred by sampling the six emissions inputs shown in Table 2. Each input is represented by a continuous random variable that can take any value in the inversion range, including the minimum and maximum values. The ranges bound the actual values used for the Diablo Canyon tracer release experiment, which are also listed in the table. The release latitude and longitude are sampled over a roughly 2 km × 2 km bounding box centered on the actual location. ~~The~~ To represent a surface release, the height of the release is varied between 1 and 10 meters above ground, with the actual height at 2 m above the surface. Potential release start times within a two hour period centered around the actual start time (08:00 local time) are considered. Similarly, ~~a range of~~ possible release durations lasting between 6 to 10 hours are tested, with the actual release occurring for 8 hours. Lastly, the inversion algorithm considered any amount between 10 and 1000 kg for the the trace gas released, with the true value at 146 kg.

The FLEXPART ensemble contains 40,000 dispersion simulations that were run and analyzed for the Diablo Canyon release. ~~The ensemble FLEXPART~~ These ensemble simulations were generated by randomly sampling both the WRF ensemble and the FLEXPART emissions variables. Random samples were drawn using a Latin hypercube design (Helton and Davis, 2003) ~~,~~ ~~which~~ assuming an 11-dimensional uniform probability distribution. Latin hypercube is a space-filling variation of Monte Carlo ~~, assuming an 11-dimensional uniform probability distribution.~~ that partitions the sampling space into N^D bins of equal probability for D dimensions and N sample points. The points are spread across this space by avoiding duplicate bin indices within each dimension. Additional discussion of Latin hypercube sampling for ensemble modeling is given in Lucas et al. (2013). To run, schedule, and manage the large FLEXPART ensemble, we utilized the Lawrence Livermore National Laboratory's *UQ Pipeline* software package (see Lucas et al., 2013). Further details on the statistical aspects of the ensemble modeling are given below in Sect. 3.

3. Statistical analysis of ensembles

3.1. Machine learning

Machine learning is used to train statistical regression functions to approximate the input-output relationships in the WRF-FLEXPART ensemble. Once trained, the machine learning functions can be evaluated very efficiently at new input values and used for uncertainty propagation, parameter estimation, Bayesian inference, and other types of statistical analysis. These functions are used for two primary purposes ~~for~~ in our work. They are used to identify and rank the effects of input features in WRF and FLEXPART on the tracer responses (i.e., a form of sensitivity analysis) and to determine the values of the inputs that yield responses that are similar to tracer observations (i.e., optimization and inverse modeling). These applications are described in more detail below.

The WRF-FLEXPART ensemble is mathematically expressed as

$$\mathbf{y} = F(\mathbf{x}_{\mathrm{WRF}}, \mathbf{x}_{\mathrm{FLX}}), \qquad (1)$$

where $\mathbf{y}$ is a vector containing information about the output or response of the simulations in the ensemble, $\mathbf{x}_{\mathrm{WRF}}$ and $\mathbf{x}_{\mathrm{FLX}}$ are vectors containing the corresponding categorical and continuous inputs to WRF and FLEXPART, respectively, and the function F represents the WRF and FLEXPART physical models. The response vector $\mathbf{y}$ is taken as either the tracer concentration at a specified location and time or a measure of the goodness-of-fit between the ensemble simulations and measurements. Complex spatiotemporal dispersion patterns are not contained in $\mathbf{y}$, although new statistical methods are being developed to capture these effects (Francom et al., 2016). Machine learning is used to approximate Eq. (1) by training on the ensemble data, which results in

$$\hat{\mathbf{y}} = F(\mathbf{x}_{\mathrm{WRF}}, \mathbf{x}_{\mathrm{FLX}}) + \epsilon_{\mathbf{y}}, \tag{2}$$

where $\hat{\mathbf{y}}$ is an approximation of $\mathbf{y}$ and $\epsilon_{\mathbf{y}}$ is the approximation error. The value of $\epsilon_{\mathbf{y}}$ is small for our analysis and is neglected for the remaining discussion (i.e., differences between $\mathbf{y}$ and $\hat{\mathbf{y}}$ are less than 10% on average, not shown). For notational convenience, the inputs $\mathbf{x}_{\mathrm{WRF}}$ and $\mathbf{x}_{\mathrm{FLX}}$ are also combined into a single input vector $\mathbf{x}$ in subsequent discussion.

We use a method called gradient boosting (GB) that fits statistical regressions to Eq. (1) using sums of decision trees following the basic notion that an individual decision tree by itself is a *weak learner*, but a combination of trees is a *strong learner* capable of fitting complex systems. The GB algorithm is briefly outlined below, and further details are available (e.g., Hastie et al., 2009). We use the GB version available in the Scikit-learn package (Pedregosa et al., 2011). GB models have been applied to studies of air quality (Carslaw and Taylor, 2009; Sayegh et al., 2016) and terrestrial ecosystems (Moisen et al., 2006; Dube et al., 2015). GB is also closely related to random forests, which we have previously used to analyze ensembles of a global climate model (Boyle et al., 2015).

As noted, a GB model is a sum of decision trees of the form

$$F(\mathbf{x}) = \sum_{m=1}^{M} T_m(\mathbf{x}; \theta), \tag{3}$$

where $T_m(\mathbf{x}; \theta)$ is the tree at stage m and θ is a set of tree fitting parameters (e.g., depth of the trees). An individual tree is described by

$$T(\mathbf{x}; \theta) = \sum_{j=1}^{J} \gamma_j I(\mathbf{x} \in R_j), \tag{4}$$

which partitions the input space $\mathbf{x}$ into J disjoint regions and assigns a value of γ_j to region R_j via the indicator function I. Starting from an initial model that fits the mean of the response, F_0, the regression model is built-up additively using a *boosting* technique that fits trees to the residuals between the current and previous stages,

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + T_m(\mathbf{x}), \tag{5}$$

where the tree output values, γ_{jm}, are defined implicitly. This expression is solved by numerically estimating gradients of a loss function (e.g. least squares) and using steepest descent optimization. A *stochastic* variation of GB is used that considers a random subset of training data during each stage, which has been shown to improve the accuracy of the fits.

Although other statistical regression methods could be used, GB offers two clear advantages for fitting the WRF-FLEXPART ensemble. First, as shown in Eq. (2), GB naturally handles heterogeneous inputs (i.e., $\mathbf{x}_{\mathrm{WRF}}$ and $\mathbf{x}_{\mathrm{FLX}}$). This makes it convenient for analyzing the combined effects of WRF inputs that vary as categories or discrete variables and FLEXPART variations that vary continuously. In addition, GB has a built-in technique for determining the influence of the inputs on the outputs during training using an established procedure called *feature importance* (e.g., see Hastie et al., 2009) . The algorithm estimates ~~feature scores for the inputs based on the positions of the nodes in the decision trees. Inputs near the top root node~~ a score for each input based on its position in each decision tree. A given decision tree is built from top-to-bottom by splitting the training data along input directions that explain most of the variance in the output. Inputs that affect the output ~~more strongly than inputs near the bottom terminal nodes, and therefore~~ strongly appear closer to the top of a decision tree and have higher feature scores. ~~Applied to the WRF-FLEXPART ensemble ,~~ Our ensemble is generated by varying inputs in WRF and FLEXPART, so the feature scores ~~are analogous to sensitivity coefficients that~~ quantify the fraction of the ensemble variance ~~caused by the changes in the inputs~~explained by these inputs. In this way, the feature scores are analogous to sensitivity coefficients used to learn about which inputs are most important in the WRF-FLEXPART ensemble.

3.2. Bayesian inverse modeling

The goal of the inverse modeling is to determine the values of the inputs to WRF and FLEXPART that provide output concentrations that best match the tracer measurements. The inversion uses an extension of our approximate Bayesian computation algorithm described in Lucas et al. (2016). The algorithm has been updated to enable joint inversions of categorical inputs in WRF and continuous inputs in FLEXPART, and to allow more flexibility in the calculation of the model-observation likelihood distance weights. In particular, the likelihood weights now utilize predictions from the GB regressions and consider more than one distance metric. Further information about the scheme is provided below and is illustrated in Fig. 4.

The inverse method applies Bayes' rule,

$$P(\mathbf{x}|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{x})P(\mathbf{x}), \tag{6}$$

to estimate $P(\mathbf{x}|\mathbf{y})$, which is the conditional probability density function (i.e., the posterior distribution) of the WRF and FLEXPART inputs, $\mathbf{x}$, given simulations and measurements of tracer concentrations, $\mathbf{y}$. The prior probability distribution of model inputs, $P(\mathbf{x})$, is an 11-dimensional uniform probability distribution over the WRF and FLEXPART random variables listed in Tables 1 and 2. As illustrated on the left hand side of Fig. 4, the prior distribution uses uniform categorical random variables for the WRF inputs and uniform continuous random variables for the FLEXPART inputs. Samples are drawn from the prior distribution using a Latin hypercube method (Helton and Davis, 2003).

The remaining term in Eq. (6) is the likelihood function, $L = P(\mathbf{y}|\mathbf{x})$, which quantifies the level of agreement between the simulated and measured tracer concentrations for a given draw from the prior distribution. Relatively high and low likelihood values correspond, respectively, to simulations that agree well and poorly with measurements. The values of model inputs that maximize the likelihood function yield peaks in the posterior distribution because of our uniform prior distribution. Moreover, variations in non-sensitive model inputs will not dramatically change the likelihood values, resulting in posterior distributions that are nearly "flat" and uniform, as in the prior distribution.

Following our previous work (Lucas et al., 2016), we compute the mean-squared-error (*mse*) between simulations and measurements as one metric for calculating high and low likelihoods. Furthermore, we include the correlation (*corr*) between observations and simulations as another metric to determine high and low likelihood values. The *corr* metric is included because it is sensitive to different aspects of model and observation differences than the *mse*. The *mse* varies with the magnitude of the differences between observations and simulations, and is expected to be mainly sensitive to changes in the source amount. The temporal correlation, on the other hand, is expected to be sensitive to changes in the arrival time and duration of the plume at the measurement locations. By combining the two metrics, *mse* and *corr*, into a single likelihood weight, we aim to constrain a larger number of input parameters than we could using either metric alone.

To account for the two metrics in the likelihood function, we use an expression of the form

$$\log L = -0.5\,[\mathbf{y_s}(\mathbf{x}) - \mathbf{y_t}]^{\mathbf{T}}\,\mathbf{\Sigma^{-1}}\,[\mathbf{y_s}(\mathbf{x}) - \mathbf{y_t}]\,, \qquad (7)$$

where $\mathbf{y_s}$ is a column vector of the metrics for a simulation at input $\mathbf{x}$,

$$\mathbf{y_s}(\mathbf{x}) = \begin{pmatrix} \mathrm{mse} \\ \mathrm{corr} \end{pmatrix}, \qquad (8)$$

$\mathbf{y_t}$ is the corresponding column vector of the metric "targets", and $\mathbf{\Sigma}$ is the 2×2 covariance matrix of model and observation errors. The highest likelihood values occur at the inputs that jointly minimize the *mse* and maximize the *corr*. Ideally, with ~~a~~ perfect model and data, the targets for *mse* and *corr* would be 0 and 1, respectively. In practice, however, models are imperfect and data is noisy, so it is usually not possible to find simulations that match the data perfectly. To avoid extrapolation, we ~~therefore~~ define the targets in $\mathbf{y_t}$ using a small number of the best fitting simulations within the ensemble, and then estimate the covariance matrix $\mathbf{\Sigma}$ using a bootstrap resampling procedure (Wilks, 2011). Further details of the target and covariance estimation are provided in Sect. 5.4.

Before computing the *mse* and *corr* in Eq. (8), the tracer concentrations are transformed using a Box-Cox power transformation with an exponent of −0.25 (see Wilks, 2011). This transformation generalizes the logarithmic transform and is used because the tracer concentrations vary over many orders of magnitude. Without it, the likelihood metrics would be skewed toward higher tracer concentrations near the release location. The Box-Cox transformation also symmetrizes the distribution of differences in Eq. (7) by removing long tails.

The Bayesian inversion is performed using GB regressions, instead of actual model simulations, to predict $\mathbf{y_s}(\mathbf{x})$ for 2 million new Latin hypercube input values. Two million points are needed to better cover the large sampling volume of the 11-dimensional prior distribution because the sampling volume varies exponentially with the number of dimensions. For instance, partitioning the ranges of only the 6 FLEXPART inputs into 10 bins each results in a volume with 10^6 bins. Running and analyzing the output of 10^6 FLEXPART simulations is computationally infeasible, so we use the ensemble of 10^4 simulations (Sect. 2) as a training dataset to build the GB regressions, and then use the regressions as surrogates for the actual models in the inversion because they can be evaluated very efficiently.

To verify the Bayesian inversion scheme, we performed a series of "synthetic data" tests using model-generated inputs and outputs. These tests are important because inverse problems often have multiple solutions and may be poorly constrained (i.e., ill-posed and ill-conditioned). Sect. 5.5 highlights the results of a synthetic data inversion test.

4. Diablo Canyon tracer release experiment

Field measurements from the Diablo Canyon nuclear power plant tracer release experiment (Thuillier, 1992) are compared to the ensemble simulations and used to test the Bayesian inversion algorithm. The Pacific Gas and Electric Company (PG&E) owns and operates Diablo Canyon and conducted the tracer experiment in 1986 to evaluate and improve PG&E's dispersion modeling capabilities in case of an accidental release. Figures 2 and 3 show the geographical setting around Diablo Canyon, which is located on the California coast in complex terrain near San Luis Obispo. The plant sits on a shelf about 26 m above sea level and is surrounded by hills with peaks about 500 m above sea level and many canyons. The hills block onshore, westerly flow, which creates challenges in simulating the effects of plumes released from Diablo Canyon on the population centers of San Luis Obispo and Pismo Beach.

PG&E conducted eight tracer release tests between August 31 and September 17, 1986. Although the large-scale wind patterns for the eight tests showed relatively similar onshore flow from the northwest, the third tracer test on September 4 experienced a strong ~~seabreeze~~ sea-breeze that presents a challenge for dispersion modeling. We therefore use the third tracer release for our inversion testing. Starting at 8 AM Pacific Daylight Time (PDT) on September 4 (15:00 UTC), 146 kg of the passive tracer sulfur hexafluoride (SF_6) was released 2-m above ground from Diablo Canyon steadily over an 8-hour period (8 AM to 4 PM PDT). The concentration of SF_6 was measured at the network of 150 tracer air sampling locations shown in Fig. 3 (black dots).

The measurement network was designed to monitor the expected tracer transport paths near terrain gaps, the entrances and exits to the inland valley, and the coastal boundary. An arc of 24 sampling sites was positioned very close to the nuclear power plant ~~release point~~release point, at a radial distance of 840 m (black dots surrounding Diablo Canyon in Fig. 3). This arc was designed to detect the initial direction of transport of released plumes and to provide details of dispersion at the nominal plant boundary. A second, linear array of 8 samplers was placed about 7 km southeast of the plant to detect transport and indicate dispersion characteristics along the anticipated principal coastal transport path for plumes released at Diablo Canyon (i.e., see site 330 in Fig. 3). Tracer sampling was done automatically by sequential pumps filling polyvinyl fluoride bags. An integrated sample was taken over each hour at each sampling site from 7 AM to 7 PM PDT. This allowed for one-hour samples prior to the tracer release for the purpose of estimating tracer background levels, and three one-hour samples after the cessation of release for the purpose of following the tracer as it traveled through the domain.

Figure 5 shows the SF_6 measurements used for the inversion. Of the 150 locations in the Diablo Canyon tracer release measurement network, 137 stations are contained in the FLEXPART domain represented by dashed rectangle in Fig. 3. The left hand side of Fig. 5 shows the pre-release SF_6 concentrations measured at these 137 stations labeled by their site identification numbers. The pre-release concentrations are used to gauge background values of SF_6. As shown, there are a handful of site ~~300~~ 3XX stations that have highly elevated values of SF_6 (above 100 ng m^{-3}). These values are well above background tropospheric SF_6 levels (Rigby et al., 2010) ~~, and are therefore~~ and are due to local pollution. ~~Sites 312–323 are affected by local pollution~~ These are sites 312–323 and are contained within the southern portion of the 840 m arc of stations surrounding Diablo Canyon. The source of the pollution is likely due to fugitive emissions from a power switchyard located within the Diablo Canyon premises. Rather than attempt to account for the extra source from the fugitive emissions, we instead exclude sites 312–323 from our analysis. After removing these sites, a dataset with 1,148 one-hour averages of the SF_6 concentration measured at 125 stations is used for the inversion (352 points are missing in the raw data). The resulting distribution of SF_6 concentrations is shown on the right hand side of Fig. 5.

5. Results

5.1. ~~Forward model plumes~~WRF ensemble with actual tracer release

~~Examples~~ Before presenting results from the large Latin hypercube ensemble, we first show dispersion results using the actual tracer release parameters with the 162 wind fields from the WRF ensemble. Figure 6 shows examples of simulated 30-minute average dispersion plumes ~~are shown in Fig. 6.~~ for two of these cases. These simulations use identical input settings and parameters in WRF and FLEXPART, except for the reanalysis fields. The plumes on the left use NARR and those on the right use ECMWF. The remaining WRF settings follow the base case values listed in Table 1, while the FLEXPART simulations use the actual values of the source release parameters listed in Table 2. These plumes therefore represent ~~the~~ our best prior knowledge in a forward modeling sense, and provide tracer concentrations that we would expect to compare reasonably well to measurements without inverse modeling.

The upper portion of the figure shows the plumes using NARR and ECMWF five hours after the release. At this stage of the simulations, there is a large spatial difference between the plumes. The dispersion using NARR is directed eastward, is spatially more confined, and does not extend downwind of Pismo Beach, as compared to the southeast directed ECMWF plume. The ECMWF plume covers a much wider region, though most of the extended area is over the ocean. Because there are not many measurement sensors over the ocean, we expect there to be smaller differences between NARR and ECMWF in the inversion algorithm than the plumes in upper part of Fig. 6 suggest.

Nine hours after the release, as shown in the lower part of the figure, the plumes using NARR and ECMWF begin to resemble each other. Both are directed to the southeast, and both have about the same spatial extent. The higher concentration area of the plumes using ECMWF are a little more dispersed near the release location (see red contour), but otherwise the differences between the two reanalysis cases are minor. ~~The next section shows that differences are substantial when the 11 model inputs are varied together.~~

5.2. ~~*Prior probability distribution of SF$_6$*~~

~~Time series of the ensemble of SF$_6$ concentrations at four representative Diablo Canyon tracer measurement sites are displayed in~~ To see the variability associated with the full WRF ensemble with the actual tracer release, Fig. ~~9. These time series , which are from the 40, 000 member Latin hypercube ensemble, provide an estimate of the prior probability distribution of SF$_6$ concentrations due to variations in meteorology and emissions because they sample the uniform random variables in WRF and FLEXPART without considering observational constraints~~7 shows the distribution of time series of SF$_6$ at the four representative locations (sites 325, 330, 338, and 413). The time series at each site displays the reference WRF-FLEXPART base simulation (black line), the tracer measurements (red squares), and different quantiles of the ~~prior SF$_6$~~ SF$_6$ distribution, including the median (solid blue line) and 5-95% range (light blue shading).

Starting with the ~~time series~~ distribution at Site 325, which is closest to the release point, the simulated SF~~$_6$~~$_6$ concentrations are negligible until about 9 AM, when the plume begins to pass over the location. The concentrations stay elevated for about 8 hours and then drop off as the trailing edge of the plume moves over the site. The ~~time series~~ distributions at the other sites show similar behavior, except that the arrival times of the plume are delayed and the peak concentrations are reduced relative to their distance from the release point. ~~Considering the large~~

Figure 7 shows that the WRF-only variations lead to fairly wide SF$_6$ distributions that span one to two orders of magnitude based on the 5-95% quantile range~~in the time series, which spans about three orders of magnitude in the SF$_6$ concentrations, the tracer measurements generally agree with the ensemble, other than during the~~. Other than the initial 2–3 ~~hours. Also note that, although the reference case appears to agree with~~ hours when background SF$_6$ was present, the figure also shows that the Diablo Canyon measurements generally fall within the distributions. Moreover, the median of the ensemble (blue line) seems to match the observations better than the ~~measurements reasonably well (i. e.~~WRF reference base case (black line).

To further examine the WRF variations, we compute the observational metrics *mse* and *corr* described in Eq. 8 for each of the 162 cases. The resulting values are displayed in Fig. 8. The best fitting WRF cases have relatively low *mse* and high *corr* and are located in the upper left portion of the figure. There appears to be a linear relationship between the metrics. WRF cases with low *mse* values also have high *corr*, and vice versa. Sect. 5.4 shows, however, that this relationship is not generally preserved when source parameter variations are included in the ensemble.

To determine the primary causes of variation in Fig. 8, the values are color-coded by reanalysis data and use different symbols for early and late time initializations (circles and squares for 9 and 15 hours before the release, respectively). Stratifying the metrics in this way shows that most of the variability is due to differences in reanalysis data. The simulations using NARR and ECMWF reanlaysis tend to fit the Diablo Canyon measurements better than CFSR. Moreover, the runs using NARR have a wider range and exhibit more variability than those using ECMWF. The initialization time also has a noticeable impact. Runs initialized 9 hours before the release (circles) have better scores than those initialized earlier (squares). Labeling the points in the figure by other WRF inputs (e.g., nudging options or PBL height) does not clearly separate the data, which suggests that additional factors are less important in optimizing the fit to the observations. Additional analysis of the effects of the WRF variations are described in the following sections (Sects. 5.3 and 5.6).

5.2. Prior probability distribution of SF$_6$

The results presented in this section and in the rest of the manuscript are based only on the 40,~~the black versus red lines) , the inversion algorithm may be able to identify a WRF configuration that provides a better match.~~000 member Latin hypercube ensemble. We exclude the 162 cases using the known release parameters that were analyzed previously in Sect. 5.1. By removing these cases from our training data, we rigorously test our Bayesian inverse method by assuming no prior knowledge about the release.

Distributions of the time series of SF$_6$ concentrations for the Latin hypercube ensemble are displayed in Fig. 9. These time series are at the same four representative locations as shown in Fig. 7. The distributions provide an estimate of the prior probability because they sample the uniform random variables in WRF and FLEXPART without considering observational constraints. The prior distributions have the same general features as the WRF-only variations in Fig. 7, but they are significantly wider. Referring to the inter-quartile range (blue dashed lines) at Site 325, for instance, the prior distribution is about two orders of magnitude larger than the corresponding distribution in Fig. 7. Similar differences between the distributions also occur at the other locations.

Comparing Figs. 7 and 9 shows that the FLEXPART source term variations account for most of the variability in the prior distribution, though it is not possible to determine which specific inputs are most influential from these figures. We therefore use gradient boosting regression in the next section to quantify the individual sources of variability in the prior distribution.

5.3. Feature scores of SF$_6$

The SF~~$_6$~~$_6$ concentrations in the prior probability distribution in Fig. 9 vary by three orders of magnitude due to variations in the inputs to WRF and FLEXPART. Gradient boosting tree regressions are used to estimate the input feature scores, where the score for a given input is analogous to the sensitivity index quantifying the fraction of the variance caused by changes in that input. The feature scores are extracted from fitting individual ~~GBTRs~~ GB regressions to the Latin hypercube ensemble at each site and for 30-minute concentration average periods. We only fit ~~GBTRs~~ GB regressions during the periods when there is significant plume ensemble variability present at all of the locations simultaneously, which occurs between 12:00 and 20:00 PM local time.

Figure 10 displays the resulting time series of the GB-based feature scores at the four representative sites. The stacked color-coded bands show the fraction of ensemble variance in the prior distribution explained by the 11 inputs, with the scores for the FLEXPART inputs at the bottom and the WRF inputs at the top of each stack. The patterns are generally similar at different locations and times, which indicates that there is not a strong spatial or temporal component to the feature scores or model sensitivities. The patterns also show that the FLEXPART and WRF inputs cumulatively account for about 80% and 20%, respectively, of the ensemble variance in Fig. 9. On the FLEXPART input side, the latitude and longitude of the release and the source amount are the most important features explaining the variance of the prior probability distribution. Among the WRF inputs, the feature scores associated with different reanalysis fields are slightly higher than the scores of the other WRF features.

Overall, the feature scores suggest that the prior uncertainty in the source term inputs are more critical than the prior uncertainty in the meteorological inputs for this particular tracer release experiment. Although the WRF inputs are not the dominant source of variability, the combined effects of the sources of meteorological uncertainty still cannot be neglected. It is also important to note that, for tracer release simulations conducted under different meteorological conditions, at different space and time scales, or that consider additional sources of uncertainty ~~,~~ or observational constraints (i.e., posterior uncertainty), the input sensitivities will likely differ from those estimated here. Moreover, the contribution of meteorological uncertainty is expected to be larger for forecast problems that are not constrained by reanalysis data.

In addition to being useful for understanding the drivers of variance in the prior probability distribution, the feature scores are also useful for interpreting the results of the Bayesian inversion in the following sections. Inputs with relatively high feature scores are often easier to constrain with observations. On this basis, therefore, we expect the posterior probability distributions for the FLEXPART ~~location~~ longitude, latitude, and source amount inputs to be relatively narrower than the other FLEXPART terms because they have the highest feature scores.

5.4. Likelihood distance metrics

Figure ~~12 displays the mean-squared-error and correlation between the ensemble simulations and the tracer measurements. The blue and red dots~~ 11 compares the 1,148 hourly-average SF$_6$ concentration measurements to two simulations in the 40,000 member ensemble. These simulations represent high and low likelihood cases (ensemble runs 32,955 7,894, respectively). The indices of the data points in the figure are arranged first by measurement site number, and then by measurement time within a given site. Low measured concentrations occur at low indices (less than about 500) and correspond to sites located upwind of the release that measured only background SF$_6$. The highest measured concentrations occur at intermediate indices (between 500 and 700) which are sites located just downwind of the release. As shown in the figure~~show the~~ , the high likelihood simulation agrees well with the measurements. It captures the location, timing, and magnitude of the actual release, resulting in a relatively low *mse* and high *corr*~~for the 40, 000 individual FLEXPART simulations, where for each simulation the metrics are computed using 1, 148 hourly-average SF~~. Not surprisingly, the FLEXPART inputs for the high likelihood case are similar to the known release values (see Table 2 and Fig. 11). The low likelihood simulation, in contrast, differs significantly from the measurements. It misses the plume at important downwind locations and predicts relatively high SF$_6$ ~~concentration measurementscollected at 125 locations within the simulation domain over the 12 hour period (352 measurements are missing). The simulations~~ concentrations at places where low concentrations were measured. The poor fit results in a relatively high *mse* and low *corr*. The low likelihood case used much less SF$_6$ than the actual amount and emitted it out over the ocean about 1.6 km to the southwest of the actual release location.

Figure 12 displays the *mse* and *corr* for all of the FLEXPART ensemble simulations. The blue and red dots show the values for the 40,000 simulations, while the yellow dots show the values for the 162 reference simulations using the known source inputs previously described and shown in Sect. 5.1 and Fig. 8. The simulations that provide a reasonable fit to the measurements have low values of *mse* and high values of *corr* and are located in the upper left quadrant of the figure, while those that disagree with the measurements are located in the lower right portion of the figure. The 50 best matching simulations are displayed using red dots in the upper left. As noted earlier, the best values of *corr* and *mse* in the figure are far from the perfect values of 1 and 0, respectively, because the WRF and FLEXPART models are imperfect and the measurements are noisy.

The points in the figure are used to estimate the terms in the likelihood function in Eq. (7) by the following procedure. First, the data points are used to form a training dataset to fit GB regressions to predict the *mse* and *corr* at new simulation input values that are not part of the ensemble. The resulting GB models fit the data very well (not shown), having coefficients of determination between actual and predicted values of $R^2 = 93\%$ and 98% for *mse* and *corr*, respectively. New Latin hypercube samples are then drawn and evaluated in the GB models to form the simulation vectors $\mathbf{y_s}(\mathbf{x})$ in Eq. (7). The remaining terms, the target vector $\mathbf{y_t}$ and covariance matrix $\mathbf{\Sigma}$, are estimated from the best fitting simulations in the figure (i.e., the red dots) by applying a bootstrap technique that resamples the tracer measurements with replacement (Wilks, 2011). We loop through the set of best fitting simulations 500 times, and each time use a random subset of the measurements (50%) to recompute *mse* and *corr*. The target vector and covariance matrix are then estimated by fitting a multivariate Gaussian distribution to the bootstrapped points, yielding the mean and standard deviation ellipses shown in the figure. The distribution is used in the inversion to calculate the posterior probability distributions of the model input values (Sect. 5.6) and tracer concentrations (Sect. 5.6).

5.5. *Inversion with synthetic data*

Before performing an inversion with the Diablo Canyon tracer data, we first apply the algorithm to "synthetic" data with known inputs and outputs as a verification test. Synthetic data is generated ~~using~~ by adding noise to the output concentrations from a randomly selected ensemble member~~as the inversion target~~. The posterior probability distribution of parameter values is computed using the previously described methods (i.e., fitting GB regressions to the *mse* and *corr* and estimating the covariance matrix as in Sect. 5.4). We draw 2 million new Latin hypercube points from the prior distribution to better cover the 11-dimensional input space, evaluate these points in the likelihood function, and compare the maximum likelihood locations to the known input values.

Figure 13 shows an example of a synthetic data inversion test using an arbitrarily selected Latin hypercube run. Other simulations have also been tested, but the results are not displayed here. The figure shows the posterior distribution of model parameter values. The FLEXPART parameters are displayed in the left hand portion of the figure using continuous distributions, while the WRF parameters are shown on the right using categorical distributions. The plots along the diagonal show univariate marginal distributions for the labeled parameters with the vertical axis indicating the normalized probability density. The off-diagonal plots show bivariate marginal distributions for the pair of parameters in the corresponding row and column with the red colors showing regions of high probability density. The known input values are denoted by the black vertical lines and circles.

For this particular test, the ~~WRF simulation~~ synthetic source was located about 850 meters to the northeast of the Diablo Canyon release that is analyzed in the next section. The WRF simulation for the synthetic test used the 06:00 UTC initialization time, the NARR reanalysis data, the MYNN TKE PBL scheme, the RUC land surface model, and no data assimilation nudging. The Bayesian inversion algorithm successfully determines these inputs, because the areas of highest posterior probability density coincide with the known values (i.e., the tallest bars and red areas overlap with the black lines and circles). All of the WRF inputs, ~~but~~ except the land surface model type, exhibit large differences across the posterior categories, which indicates that the inputs are well constrained by the data and metrics. In particular, there is little to no posterior weight associated with the other initialization time and reanalysis fields. The relatively small differences across the LSM categories are thought to occur because the plume is predominantly transported over the ocean and sampled near the coast (see Figs. 3 and 6), and therefore the *mse* and *corr* metrics are not sensitive to changes in the land surface model.

~~The~~ Table 3 compares the positions of the peak values in the posterior distribution for the FLEXPART dispersion inputs ~~also~~ to the actual input values. The maximum likelihood values agree exceedingly well with the ~~known~~ actual input values. Except for the release altitude, the algorithm infers the location, amount, and timing of the source. As determined from the widths of the posteriors, the release latitude and longitude are the best constrained FLEXPART inputs, followed by the source amount and duration. The inferred location lies only about 50 meters away from the actual value. The posterior distribution for the release altitude is relatively unchanged from the flat prior distribution, ~~suggesting that the data and metrics are not sufficient to constrain~~ which is not surprising because FLEXPART is insensitive to the relatively small variations of the release altitude (0 to 10 meters). As previously noted, there is a reasonable correspondence between the widths of the FLEXPART posterior widths in Fig. 13 and the size of the feature scores in Fig. 10.

These results, along with other synthetic data tests that are not shown, provide confidence that the inversion algorithm appears to be functioning adequately. The algorithm returns values for the WRF and FLEXPART inputs that are close to the actual values for most of the parameters. For the release height, we are also satisfied with the non-informative values provided by the algorithm because we expected a relatively flat posterior distribution. In future work, we will broaden the range of release heights to test the algorithm for elevated and surface releases.

5.6. Inversion with Diablo Canyon tracer data

For the inversion using the SF_6 measurements, we draw another 2 million Latin hypercube points from the prior distribution for WRF and FLEXPART, evaluate the points in the GB fits for *mse* and *corr*, and then compute the likelihood weights relative to the target and covariance displayed in Fig. 12. The resulting posterior distribution of WRF and FLEXPART parameters is shown in Fig. 14.

As is the case with the synthetic inversion tests, the actual values for the location, start, duration, and amount of the SF_6 release are known for this tracer experiment. The inversion, however, assumes that the release parameters are unknown and uses the measurements to infer their values. ~~As shown in the figure~~The inversion results are compared to the actual values in Table 3 and Fig. 14. As shown, the data and algorithm are sufficient to determine most of the FLEXPART source term parameters, because the ~~positions of the~~ maximum likelihood values of the parameters closely match the known experimental values~~(black lines and circles in Fig. 14).~~. The close agreement between the two implies that the WRF and FLEXPART models do not have any severe deficiencies that prevent them from accurately simulating tracer transport for this experiment. The synthetic data inversion tests from the previous section would not expose model deficiencies because the same deficiency would be present in both the simulations and target, and hence would be subtracted out of the analysis.

Referring to Fig. 14, we see that the marginal distributions for the latitude, longitude, and amount of the release have the sharpest peaks and are therefore the most constrained by the measurements. The source start and duration are also moderately constrained, though the distribution for release height is unconstrained and remains essentially flat. The posterior distribution also suggests that the release duration lasts longer than the 8 hour period used in the experiment. ~~It~~ To some extent, the one-hour average observations used in the inversion limit the constraints on the start time and duration. Moreover, it is difficult to simulate rapid changes associated with the leading and trailing edges of the plume, and so the model and methods may be smoothing out these features and causing the overestimation of the duration. Other likelihood metrics besides *mse* and *corr* may help alleviate this issue.

The posterior distribution also shows a strong covariance relationship between the release latitude and longitude (see bivariate distribution in the upper left of Fig. 14). An area of relatively high probability stretches from northwest of the actual release location to the southeast. The shape of this covariance stems from the large-scale flow pattern and nearby measurement locations. The general direction of the flow for the release period is from the northwest to the southeast, and the release point is situated within a fairly close arc of sensors (see Fig. 3). As long as the release stays within this arc, moving the release location slightly upwind or downwind will not greatly affect the simulated concentrations at the sensor locations. If the release location is moved orthogonal to the flow or outside of the arc, however, FLEXPART will simulate SF_6 at sensors where none was measured, and vice versa.

The WRF configurations in the posterior distribution are displayed on the right hand side of Fig. 14. Unlike the known FLEXPART inputs described above or the known inputs in the synthetic data experiments, the actual values of the meteorology are not known here (i.e., there are no black lines or circles for WRF). The configurations that minimize the *mse* and maximize the *corr* with the SF_6 measurements are represented by the tallest bars in the univariate distributions along the diagonal and by the red-colored squares or bands in the off-diagonal bivariate distributions.

As shown in the figure, the maximum likelihood configuration consists of the 06:00 UTC initialization time, the ECMWF reanalysis fields, the YSU PBL scheme, the RUC land surface model, and no data assimilation nudging. Some of these configuration settings may, at first, seem surprising. For example, the NARR reanalysis fields have a higher spatial resolution than the ECMWF fields and therefore may be expected to perform better. Likewise, the option to run without data assimilation seems to outperform the options with assimilation. Referring to the figure for these cases, the posterior distribution still has significant probability density for both the NARR and low nudging options. Compared to the posterior distribution in the synthetic data inversion, the WRF inputs are not as strongly constrained using the tracer data, especially the inputs for the land surface model and nudging. Only two of the WRF inputs have settings with negligible probability, the earlier initialization time and CFSR reanalysis data. The alternate PBL schemes also have relatively low probability. We therefore conclude that the winds generated using the 06:00 UTC initialization time, the NARR or ECMWF reanalysis fields, and the YSU PBL scheme will optimize our likelihood metrics, and that there is not a preferred land surface model or nudging option.

5.7. *Posterior probability distribution of SF_6*

The ensemble time series in Fig. 9 are based on sampling the prior distribution of input parameters, which results in a substantial spread of SF_6 concentrations over two to three orders of magnitude. Most of these ensemble members do not agree well with the tracer measurements, so we estimate the posterior distribution of SF_6 concentrations by applying the likelihood weights of Sect. 5.4 and 5.6 to the ensemble time series in Fig. 9.

Figure 15 displays the posterior distribution of the time series of SF_6 concentrations. As before, the time series show the tracer measurements (red squares) and quantiles of the posterior distribution, including the median (solid blue line), first and third quartiles (dashed blue lines), and 5-95% range (light blue area). The general features of the posterior distribution are similar to the prior distribution, except that the ensemble spread has been greatly reduced and the quantiles have shifted to higher values.

Other than at site 325, where a large spread remains, the 5-95% range covers about one order of magnitude, or a reduction of two orders of magnitude. Even with the reduced range, most of the measurements still fall within the light blue area. We do not expect all of the measurements to lie in the 5-95% range because the likelihood metrics consider the aggregation of all of the sites and times. In order to achieve an overall higher likelihood, individual measurement points may be farther away from the median in the posterior distribution than they were in the prior.

In addition to the reduction in the variance, Fig. 15 also shows that the median and other quantiles of the posterior distribution shift to higher concentrations. This shift occurs because the source term parameter variations in the prior distribution lead to many simulations having SF_6 concentrations that are too low relative to the measurements. The likelihood weights discount these simulations,

6. Conclusions

We have developed an ensemble-based Bayesian inverse modeling system that can determine information about an atmospheric release from a nuclear power plant using measurements collected a relatively safe distance downwind from the plant. The system uses an ensemble of WRF simulations to capture uncertainty in meteorological fields and an ensemble of FLEXPART dispersion simulations to vary factors related to emissions.

Machine learning algorithms are trained on the input-output relationships in the meteorological and dispersion ensemble, resulting in statistical surrogate models that mimic the behavior of the actual WRF and FLEXPART models, but that can be evaluated very rapidly at millions of new input value combinations.

Using our system, we can determine the input factors that are most important for understanding and reducing uncertainty in the ensemble (i.e., sensitivity analysis) and can estimate the values of the model inputs that provide likely matches between model output and field measurements (i.e., inverse modeling). Bayes' rule is used for the inversion, which provides probability distribution functions of model inputs and outputs constrained by observations and that serve as a quantitative assessment of model performance. The inversion is designed to estimate the location, timing, and amount of material released to the atmosphere, and to determine the best categories of settings for running a meteorological model. The inversion system ~~can~~ should be able to handle, without difficulty, additional factors related to the transport and dispersion of potential materials released during a nuclear power plant accident (e.g., wet and dry deposition of soluble radioactive products).

Our ensemble system is tested against a tracer release experiment conducted near the Diablo Canyon nuclear power plant located in the rugged terrain of coastal California (Thuillier, 1992). An ensemble of 40,000 dispersion simulations is created using a Monte Carlo method to sample uncertainty in 6 source term parameters in FLEXPART and 5 meteorological categories distributed among 162 unique configurations in WRF. The variance of the resulting unconstrained tracer concentration ensemble is substantial (i.e., the prior distribution), covering a 5% to 95% concentration probability range of about four orders of magnitude. About 80% of the unconstrained prior variance is due to source term parameter variations, with about half the overall variance coming from just three input parameters (release amount, latitude, and longitude). Although the meteorological inputs are not dominant sources of ensemble variability, they cumulatively account for 20% of the variance in the prior distribution and are important because their uncertainty is not easily reduced.

By calculating the mean squared error and correlation between the tracer measurements and the surrogate model predictions, the Bayesian inversion algorithm produces a posterior distribution of model inputs and outputs for the tracer release experiment. Even though the source term parameters are initially unknown in the inversion (i.e., ~~an~~ we used a non-informative prior), the most likely posterior values of the FLEXPART inputs closely estimate the actual values used in the tracer release experiment, which demonstrates a successful inversion. Table 3 summarizes the results of the tracer release source inversion. As shown, the most likely values of inversion algorithm are within about 200 meters of the release location, within 5 minutes and 50 minutes of the starting time and duration, respectively, and within 22% of the actual release amount. Furthermore, the posterior values of the WRF inputs show a preference for particular configurations involving the later initialization time, YSU PBL scheme, and NARR and ECMWF reanalysis fields. Compared to the large concentration spread in the prior distribution, the posterior variance of tracer concentrations is greatly reduced and better tracks the measurements, thus indicating a good correspondence between the posterior inputs and outputs.

It is important to keep in mind that the ensemble and inversion methods can be applied to problems other than nuclear power plant releases. While the location of a nuclear power plant release is generally restricted to reactor buildings or other nearby facilities, the inversion algorithm can also determine an arbitrary release location from within a large area (e.g., 100's of km^2) if suitable observations are available.

The number of measurements can also affect the quality of the inversion. The Diablo Canyon field experiment had a large number of sensors to measure SF_6, and we used 1,148 data points to constrain 11 model parameters. Fewer measurements would be available for most real-world events, and so in future work we plan to quantify the impacts of measurement density on the inversion. In preliminary analysis for Diablo Canyon, we found that using only about 200 randomly selected measurement points still resulted in GB regressions accurate enough for estimating influential source parameters. Because some measurements are more important than others, we can also utilize network optimization techniques (e.g., Lucas et al., 2015) to further reduce the number of measurements, as needed to apply our method to compliance monitoring of the nuclear test ban treaty (Issartel and Baverel, 2003), emissions from large-scale industrial accidents (e.g., Heng et al., 2016), and other real-world applications.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was funded by Laboratory Directed

Research and Development at LLNL under project tracking code PLS-14ERD006. The manuscript is released under UCRL number LLNL-JRNL-710162. The authors thank PG&E for access to the Diablo Canyon measurement data, and Livermore Computing for providing computational resources through an institutional allocation for the Monte Carlo simulations. The authors also thank Devin Francom and Bruno Sansó from UC Santa Cruz, and Vera Bulaevskaya from LLNL for invaluable discussions about the statistical analysis.

References

Albergel, A., Martin, D., Strauss, B., Gros, J.M., 1988. The Chernobyl accident: Modelling of dispersion over europe of the radioactive plume and comparison with air activity measurements. Atmospheric Environment 22, 2431 – 2444. doi:10.1016/0004-6981(88)90475-1.

An, X., Yao, B., Li, Y., Li, N., Zhou, L., 2014. Tracking source area of Shangdianzi station using Lagrangian particle dispersion model of FLEXPART. Meteorological Applications 21, 466–473. doi:10.1002/met.1358.

Andreev, I., Hittenberger, M., Hofer, P., Kromp-Kolb, H., Kromp, W., Seibert, P., Wotawa, G., 1998. Risks due to beyond design base accidents of nuclear power plants in Europe the methodology of riskmap. Journal of Hazardous Materials 61, 257 – 262. doi:http://dx.doi.org/10.1016/S0304-3894(98)00130-7.

Athey, G.F., Fosmire, C., Mohseni, A., Ramsdell, J., Sjoreen, A., 1999. Radiological Assessment System for Consequence Analysis (RASCAL) Version 3.0. American Nuclear Society.

Avey, L., Garrett, T.J., Stohl, A., 2007. Evaluation of the aerosol indirect effect using satellite, tracer transport model, and aircraft data from the International Consortium for Atmospheric Research on Transport and Transformation. Journal of Geophysical Research: Atmospheres 112. doi:10.1029/2006JD007581. d10S33.

Becker, A., Wotawa, G., Ringbom, A., Saey, P.R.J., 2010. Backtracking of noble gas measurements taken in the aftermath of the announced october 2006 event in North Korea by means of PTS methods in nuclear source estimation and reconstruction. Pure and Applied Geophysics 167, 581–599. doi:10.1007/s00024-009-0025-0.

Benjamin, S.G., Grell, G.A., Brown, J.M., Smirnova, T.G., Bleck, R., 2004. Mesoscale weather prediction with the RUC hybrid isentropic terrain-following coordinate model. Monthly Weather Review 132, 473–494. doi:10.1175/1520-0493(2004)132<0473:MWPWTR>2.0.CO;2.

Berner, J., Ha, S.Y., Hacker, J.P., Fournier, A., Snyder, C., 2011. Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. Monthly Weather Review 139, 1972–1995. doi:10.1175/2010MWR3595.1.

Boyle, J.S., Klein, S.A., Lucas, D.D., Ma, H.Y., Tannahill, J., Xie, S., 2015. The parametric sensitivity of CAM5's MJO. Journal of Geophysical Research: Atmospheres 120, 1424–1444. doi:10.1002/2014JD022507.

Brioude, J., Arnold, D., Stohl, A., Cassiani, M., Morton, D., Seibert, P., Angevine, W., Evan, S., Dingwell, A., Fast, J.D., Easter, R.C., Pisso, I., Burkhart, J., Wotawa, G., 2013a. The lagrangian particle dispersion model flexpart-wrf version 3.1. Geoscientific Model Development 6, 1889–1904. URL: http://www.geosci-model-dev.net/6/1889/2013/, doi:10.5194/gmd-6-1889-2013.

Brioude, J., Arnold, D., Stohl, A., Cassiani, M., Morton, D., Seibert, P., Angevine, W., Evan, S., Dingwell, A., Fast, J.D., Easter, R.C., Pisso, I., Burkhart, J., Wotawa, G., 2013b. The lagrangian particle dispersion model flexpart-wrf version 3.1. Geoscientific Model Development 6, 1889–1904. URL: http://www.geosci-model-dev.net/6/1889/2013/, doi:10.5194/gmd-6-1889-2013.

Carslaw, D.C., Taylor, P.J., 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. Atmospheric Environment 43, 3563–3570. doi:10.1016/j.atmosenv.2009.04.001.

Chino, M., Nakayama, H., Nagai, H., Terada, H., Katata, G., Yamazawa, H., 2011. Preliminary estimation of release amounts of 131I and 137Cs accidentally discharged from the Fukushima Daiichi Nuclear Power Plant into the atmosphere. Journal of Nuclear Science and Technology 48, 1129–1134. doi:10.1080/18811248.2011.9711799.

Christoudias, T., Proestos, Y., Lelieveld, J., 2014. Global risk from the atmospheric dispersion of radionuclides by nuclear power plant accidents in the coming decades. Atmospheric Chemistry and Physics 14, 4607–4616. URL: http://www.atmos-chem-phys.net/14/4607/2014/, doi:10.5194/acp-14-4607-2014.

Davoine, X., Bocquet, M., 2007. Inverse modelling-based reconstruction of the chernobyl source term available for long-range transport. Atmospheric Chemistry and Physics 7, 1549–1564. URL: http://www.atmos-chem-phys.net/7/1549/2007/, doi:10.5194/acp-7-1549-2007.

Dube, T., Mutanga, O., Abdel-Rahman, E.M., Ismail, R., Slotow, R., 2015. Predicting eucalyptus spp. stand volume in zululand, south africa: an analysis using a stochastic gradient boosting regression ensemble with multi-source data sets. International Journal of Remote Sensing 36, 3751–3772. doi:10.1018/01431161.2015.1070316.

Duhdia, J., 1996. A multi-layer soil temperature model for MM5. Sixth PSU/NCAR Mesoscale Model Users' Workshop.

Ek, M.B., Mitchell, K.E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., Tarpley, J.D., 2003. Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model. Journal of Geophysical Research: Atmospheres 108, n/a–n/a. URL: http://dx.doi.org/10.1029/2002JD003296, doi:10.1029/2002JD003296. 8851.

Forster, C., Stohl, A., Seibert, P., 2007. Parameterization of convective transport in a lagrangian particle dispersion model and its evaluation. Journal of Applied Meteorology and Climatology 46, 403–422. doi:10.1175/JAM2470.1.

Francom, D., Sanso, B., Bulaevskaya, V., Lucas, D., 2016. Statistical surrogate modeling of atmospheric dispersion events using Bayesian adaptive splines. EOS Transactions AGU, Fall Meeting Supplement, Abstract GC31F-1162.

Gudiksen, P.H., Harvey, T.F., Lange, R., 1989. Chernobyl source term, atmospheric dispersion, and dose estimation. Health Physics 57, 697–706.

Hanna, S.R., 1982. Applications in Air Pollution Modeling. Springer Netherlands, Dordrecht. pp. 275–310. doi:10.1007/978-94-010-9112-1_7.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics, New York, NY, USA. doi:10.1007/b94608.

Helton, J.C., Davis, F.J., 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering & System Safety 81, 23 – 69.

Heng, Y., Hoffmann, L., Griessbach, S., Rößler, T., Stein, O., 2016. Inverse transport modeling of volcanic sulfur dioxide emissions using large-scale simulations. Geoscientific Model Development 9, 1627–1645. doi:10.5194/gmd-9-1627-2016.

Hersbach, H., Peubey, C., Simmons, A., Berrisford, P., Poli, P., Dee, D., 2015. ERA-20CM: a twentieth-century atmospheric model ensemble. Quarterly Journal of the Royal Meteorological Society 141, 2350–2375. doi:10.1002/qj.2528.

Hertel, O., Christensen, J., Runge, E.H., Asman, W.A., Berkowicz, R., Hovmand, M.F., Hov, O., 1995. Development and testing of a new variable scale air pollution model ACDEP. Atmospheric Environment 29, 1267 – 1290. doi:10.1016/1352-2310(95)00067-9.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., , Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing 81, 345–354.

Hong, S.Y., Noh, Y., Dudhia, J., 2006. A new vertical diffusion package with an explicit treatment of entrainment processes. Monthly Weather Review 134, 2318–2341. doi:10.1175/MWR3199.1.

Issartel, J.P., Baverel, J., 2003. Inverse transport for the verification of the comprehensive nuclear test ban treaty. Atmospheric Chemistry and Physics 3, 475–486. doi:10.5194/acp-3-475-2003.

Janjić, Z.I., 1994. The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. Monthly Weather Review 122, 927–945. doi:10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Katata, G., Chino, M., Kobayashi, T., Terada, H., Ota, M., Nagai, H., Kajino, M., Draxler, R., Hort, M.C., Malo, A., Torii, T., Sanada, Y., 2015. Detailed source term estimation of the atmospheric release for the fukushima daiichi nuclear power station accident by coupling simulations of an atmospheric dispersion model with an improved deposition scheme and oceanic dispersion model. Atmospheric Chemistry and Physics 15, 1029–1070. doi:10.5194/acp-15-1029-2015.

Lauritzen, B., Mikkelsen, T., 1999. A probabilistic dispersion model applied to the long-range transport of radionuclides from the Chernobyl accident. Atmospheric Environment 33, 3271 – 3279.

Lee, J.A., Kolczynski, W.C., McCandless, T.C., Haupt, S.E., 2012. An objective methodology for configuring and down-selecting an nwp ensemble for low-level wind prediction. Monthly Weather Review 140, 2270–2286. doi:10.1175/MWR-D-11-00065.1.

Liu, Y., Bourgeois, A., Warner, T., Swerdlin, S., Hacker, J., 2005. An implementation of observation-nudging based on FDDA into WRF for supporting AFEC test operations. Paper 10.7. 2005 WRF User Workshop.

Liu, Y., Warner, T., Wu, W., Roux, G., Cheng, W., Liu, Y., Chen, F., Delle Monache, L., Mahoney, W., Swerdlin, S., 2009. A versatile WRF and MM5-based weather analysis and forecasting system for supporting wind energy prediction. Paper 17B.3. URL: http://n2t.net/ark:/85065/d71n8075. 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction.

Lucas, D.D., Gowardhan, A., Cameron-Smith, P., Baskett, R.L., 2016. Impact of meteorological inflow uncertainty on tracer transport and source estimation in urban atmospheres. Atmospheric Environment 143, 120–132. doi:10.1016/j.atmosenv.2016.08.019.

Lucas, D.D., Klein, R., Tannahill, J., Ivanova, D., Brandon, S., Domyancic, D., Zhang, Y., 2013. Failure analysis of parameter-induced simulation crashes in climate models. Geoscientific Model Development 6, 1157–1171. doi:10.5194/gmd-6-1157-2013.

Lucas, D.D., Yver-Kwok, C., Cameron-Smith, P., Graven, H., Bergmann, D., Guilderson, T.P., Weiss, R., Keeling, R., 2015. Designing optimal greenhouse gas observing networks that consider performance and cost. Geoscientific Instrumentation, Methods and Data Systems 4, 121–137. URL: http://dx.doi.org/10.5194/gi-4-121-2015, doi:10.5194/gi-4-121-2015.

Lundquist, K.A., Chow, F.K., Lundquist, J.K., 2010. An Immersed Boundary Method for the Weather Research and Forecasting Model. Monthly Weather Review 138, 796–817. doi:10.1175/2009MWR2990.1.

McMahon, T.A., Denison, P.J., 1979. Empirical atmospheric deposition parameters - a survey. Atmospheric Environment 13, 571–585. doi:10.1016/0004-6981(79)90186-0.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., Shi, W., 2006. North American Regional Reanalysis. Bulletin of the American Meteorological Society 87, 343–360. doi:10.1175/BAMS-87-3-343.

Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E., Edwards Jr., T.C., 2006. Predicting tree species presence and basal area in utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based models. Ecological Modelling 199, 176–187. doi:10.1016/j.ecolmodel.2006.05.021.

Mullen, S.L., Baumhefner, D.P., 1994. Monte Carlo simulations of explosive cyclogenesis. Monthly Weather Review 122, 1548–1567. doi:10.1175/1520-0493(1994)122<1548:MCSOEC>2.0.CO;2.

Nakanishi, M., Niino, H., 2006. An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. Boundary-Layer Meteorology 119, 397–407. doi:10.1007/s10546-005-9030-8.

Nasstrom, J.S., Sugiyama, G., Baskett, R.L., Larsen, S.C., Bradley, M.M., 2007. The National Atmospheric Release Advisory Center modelling and decision-support system for radiological and nuclear emergency preparedness and response. International Journal of Emergency Management 4, 524 – 550.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pöllänen, R., Valkama, I., Toivonen, H., 1997. Transport of radioactive particles from the chernobyl accident. Atmospheric Environment 31, 3575 – 3590.

Rigby, M., Mühle, J., Miller, B.R., Prinn, R.G., Krummel, P.B., Steele, L.P., Fraser, P.J., Salameh, P.K., Harth, C.M., Weiss, R.F., Greally, B.R., O'Doherty, S., Simmonds, P.G., Vollmer, M.K., Reimann, S., Kim, J., Kim, K.R., Wang, H.J., Olivier, J.G.J., Dlugokencky, E.J., Dutton, G. S. a nd Hall, B.D., Elkins, J.W., 2010. History of atmospheric sf¡sub¿6¡/sub¿ from 1973 to 2008. Atmospheric Chemistry and Physics 10, 10305–10320. URL: http://www.atmos-chem-phys.net/10/10305/2010/, doi:10.5194/acp-10-10305-2010.

Ryall, D., Maryon, R., 1998. Validation of the UK Met. Office's NAME model against the ETEX dataset. Atmospheric Environment 32, 4265 – 4276. doi:10.1016/S1352-2310(98)00177-0.

Saha, S., Moorthi, S., Pan, H.L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.T., Chuang, H.Y., Juang, H.M.H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Delst, P.V., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Dool, H.V.D., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R.W., Rutledge, G., Goldberg, M., 2010. The NCEP Climate Forecast System Reanalysis. Bulletin of the American Meteorological Society 91, 1015–1057. doi:10.1175/2010BAMS3001.1.

Sayegh, A., Tate, J.E., Ropkins, K., 2016. Understanding how roadside concentrations of no$_x$ are influenced by the background levels, traffic density, and meteorological conditions using Boosted Regression Trees. Atmospheric Environment 127, 163–175. doi:10.1016/j.atmosenv.2015.12.024.

Seinfeld, J., Pandis, S., 2006. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. A Wiley-Interscience publication, Wiley.

Skamarock, W.C., Klemp, J.B., 2007. A time-split nonhydrostatic atmospheric model for research and NWP applications. J. Comp. Phys. 227, 3465–3485.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M.G., Huang, X.Y., Wang, W., Powers, J.G., 2008. A description of the Advanced Research WRF Version 3. Technical Report NCAR TN-475+STR. NCAR. URL: http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.

Slinn, W.G.N., 1982. Predictions for particle deposition to vegetative canopies. Atmospheric Environment (1967) 16, 1785 – 1794. doi:10.1016/0004-6981(82)90271-2.

Stauffer, D.R., Seaman, N.L., 1994. Multiscale four-dimensional data assimilation. Journal of Applied Meteorology 33, 416–434. doi:10.1175/1520-0450(1994)033<0416:MFDDA>2.0.CO;2.

Stensrud, D.J., Bao, J.W., Warner, T.T., 2000. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. Monthly Weather Review 128, 2077–2107. URL: http://dx.doi.org/10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2, doi:10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

Stohl, A., Forster, C., Frank, A., Seibert, P., Wotawa, G., 2005. Technical note: The lagrangian particle dispersion model flexpart version 6.2. Atmospheric Chemistry and Physics 5, 2461–2474. URL: http://www.atmos-chem-phys.net/5/2461/2005/, doi:10.5194/acp-5-2461-2005.

Stohl, A., Hittenberger, M., Wotawa, G., 1998. Validation of the Lagrangian particle dispersion model {FLEXPART} against large-scale tracer experiment data. Atmospheric Environment 32, 4245 – 4264. doi:10.1016/S1352-2310(98)00184-8.

Stohl, A., Prata, A.J., Eckhardt, S., Clarisse, L., Durant, A., Henne, S., Kristiansen, N.I., Minikin, A., Schumann, U., Seibert, P., Stebel, K., Thomas, H.E., Thorsteinsson, T., Tørseth, K., Weinzierl, B., 2011. Determination of time- and height-resolved volcanic ash emissions and their use for quantitative ash dispersion modeling: the 2010 Eyjafjallajökull eruption. Atmospheric Chemistry and Physics 11, 4333–4351. URL: http://www.atmos-chem-phys.net/11/4333/2011/, doi:10.5194/acp-11-4333-2011.

Stohl, A., Seibert, P., Wotawa, G., Arnold, D., Burkhart, J.F., Eckhardt, S., Tapia, C., Vargas, A., Yasunari, T.J., 2012. Xenon-133 and caesium-137 releases into the atmosphere from the fukushima dai-ichi nuclear power plant: determination of the source term, atmospheric dispersion, and deposition. Atmospheric Chemistry and Physics 12, 2313–2343. URL: http://www.atmos-chem-phys.net/12/2313/2012/, doi:10.5194/acp-12-2313-2012.

Stohl, A., Thomson, D.J., 1999. A density correction for Lagrangian particle dispersion models. Boundary-Layer Meteorology 90, 155–167. doi:10.1023/A:1001741110696.

Suh, K.S., Han, M.H., Jung, S.H., Lee, C.W., 2009. Numerical simulation for a long-range dispersion of a pollutant using Chernobyl data. Mathematical and Computer Modelling 49, 337 – 343.

Terada, H., Katata, G., Chino, M., Nagai, H., 2012. Atmospheric discharge and dispersion of radionuclides during the Fukushima Dai-ichi Nuclear Power Plant accident. Part II: verification of the source term and analysis of regional-scale atmospheric dispersion. Journal of Environmental Radioactivity 112, 141–154. doi:10.1016/j.jenvrad.2012.05.023.

Thuillier, R.H., 1992. Evaluation of a puff dispersion model in complex terrain. Journal of the Air & Waste Management Association 42, 290–297. doi:10.1080/10473289.1992.10466992.

Wahlen, M., Kunz, C., Matuszek, J., Mahoney, W., Thompson, R., 1980. Radioactive plume from the Three Mile Island accident: xenon-133 in air at a distance of 375 kilometers. Science 207, 639–640. doi:10.1126/science.7352276.

Wesely, M.L., Hicks, B.B., 1977. Some factors that affect the deposition rates of sulfur dioxide and similar gases on vegetation. Journal of the Air Pollution Control Association 27, 1110–1116. doi:10.1080/00022470.1977.10470534.

Wilks, D., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press, Academic Press.

Wotawa, G., Becker, A., Kalinowski, M., Saey, P., Tuma, M., Zähringer, M., 2010. Computation and analysis of the global distribution of the radioxenon isotope 133xe based on emissions from nuclear power plants and radioisotope production facilities and its relevance for the verification of the nuclear-test-ban treaty. Pure and Applied Geophysics 167, 541–557. doi:10.1007/s00024-009-0033-0.

Zheng, X., Chen, Z., 2011. Inverse calculation approaches for source determination in hazardous chemical releases. Journal of Loss Prevention in the Process Industries 24, 293 – 301. doi:10.1016/j.jlp.2011.01.002.

Table 1: Categorical Random Variables for WRF Ensemble.

Categorical Variable	Description	Category	Label
1. WRF_initim	Initialization time	1986-09-04-06	0*
		1986-09-04-00	1
2. WRF_reanalysis	Reanalysis data	NARR	0*
		ECMWF	1
		CFSR	2
3. WRF_nudge	Nudging	Off	0
		Low	1*
		High	2
4. WRF_pbl	Boundary layer physics	YSU	0
		MYJ TKE	1*
		MYNN TKE	2
5. WRF_lsm	Land surface model	Thermal diffusion	0
		Noah	1*
		RUC	2

∗ denotes a WRF setting used for the reference base simulation.

Table 2: Continuous Random Variables for FLEXPART Tracer Ensemble.

Continuous Variable	Description	Actual Value	Inversion Range
1. FLX_loc_lat	Release latitude	35.2111°N	[35.1977, 35.2250]
2. FLX_loc_lon	Release longitude	120.8543°W	[120.8708, 120.8384]
3. FLX_zlev_bot	Release altitude	2 meters	[1, 10]
4. FLX_source_start	Release start	08:00 local time	[07:00, 09:00]
5. FLX_source_duration	Release duration	8 hours	[6, 10]
6. FLX_source_amount	Release magnitude	146.016 kg	[10, 1000]

Table 3: ~~Most Likely~~ Inversion Estimates Versus Actual Source Parameters~~for the Diablo Canyon Tracer Release~~.

Source Term	~~Maximum Likelihood~~ Synthetic		Diablo Canyon	
	Estimated	Actual	Estimated	Actual
Latitude	35.2170°N	35.2165°N	35.2125°N	35.2111°N
Longitude	120.8478°W	120.8475°W	120.8560°W	120.8543°W
Start time	08:35	08:43	08:05 ~~local time~~	08:00 ~~local time~~
Duration	9.17 hours	8.75 hours	8.83 hours	8 hours
Amount	121.153 kg	95.54 kg	177.830 kg	146.016 kg

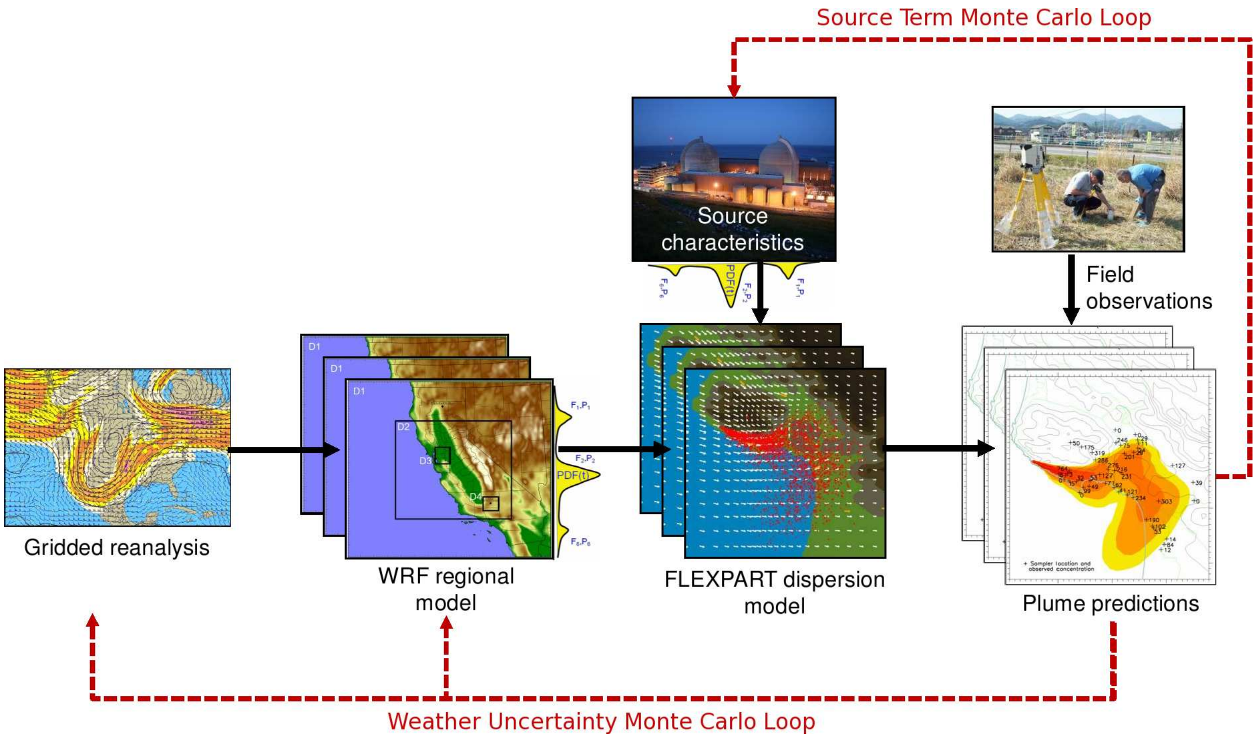


Figure 1: ~~Overview~~ An overview of ~~the~~ our ensemble ~~WRF-FLEXPART~~ weather and dispersion system for inverse modeling and ~~quantifying~~ uncertainty applications. The system uses WRF to calculate wind fields, and FLEXPART to transport materials in the atmosphere. Information about atmospheric releases is inferred by minimizing differences between plume predictions ~~due to uncertainty in meteorology~~ and ~~source release characteristics. The red dashed arrows show~~ field observations through Monte Carlo sampling loops ~~used to improve the differences between simulations~~ (red dashed arrows). The loops sample different reanalysis datasets and ~~observations~~model configurations in WRF and source term inputs in FLEXPART.

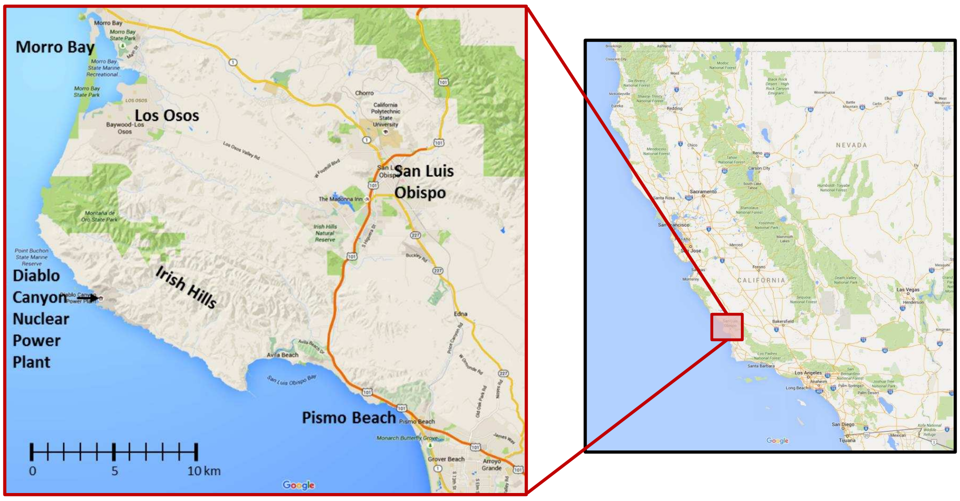


Figure 2: Maps showing the geographical location of the Diablo Canyon nuclear power plant on the coast of central California, near San Luis Obispo and Pismo Beach.

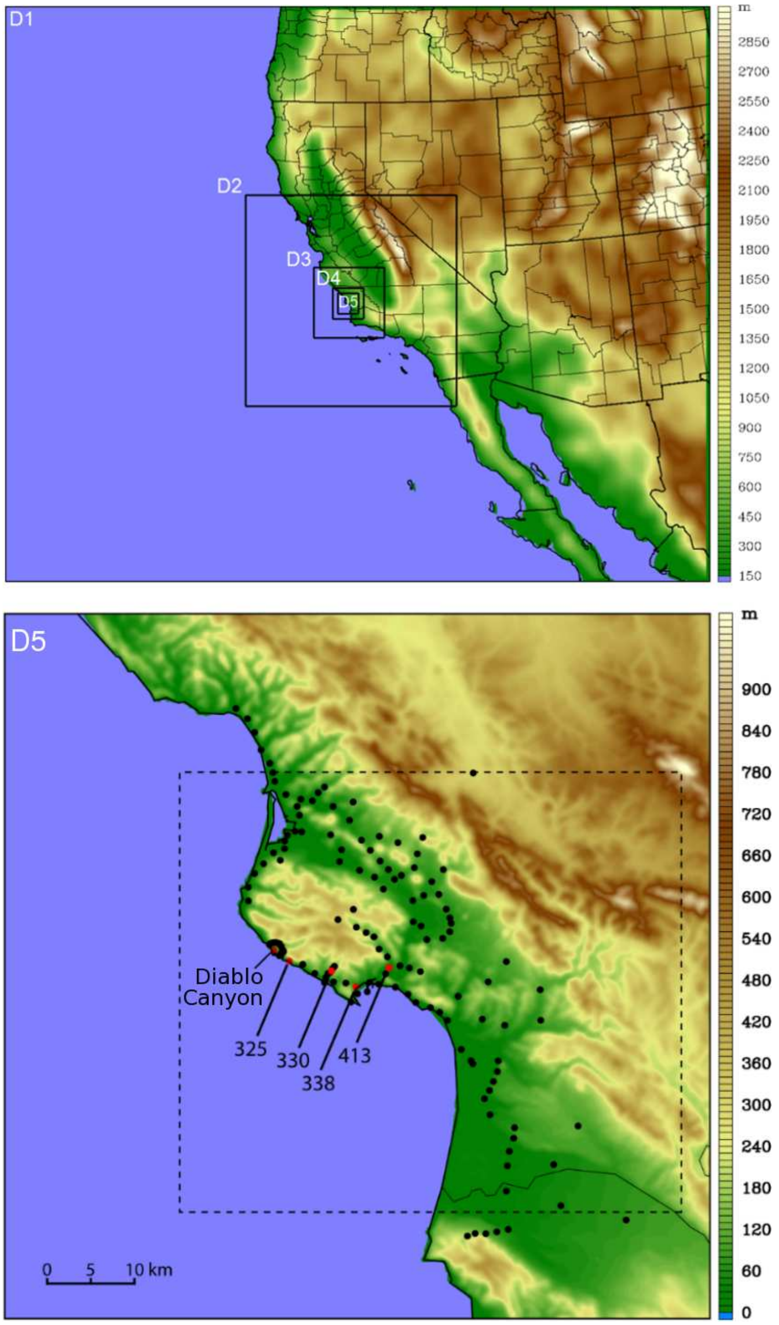
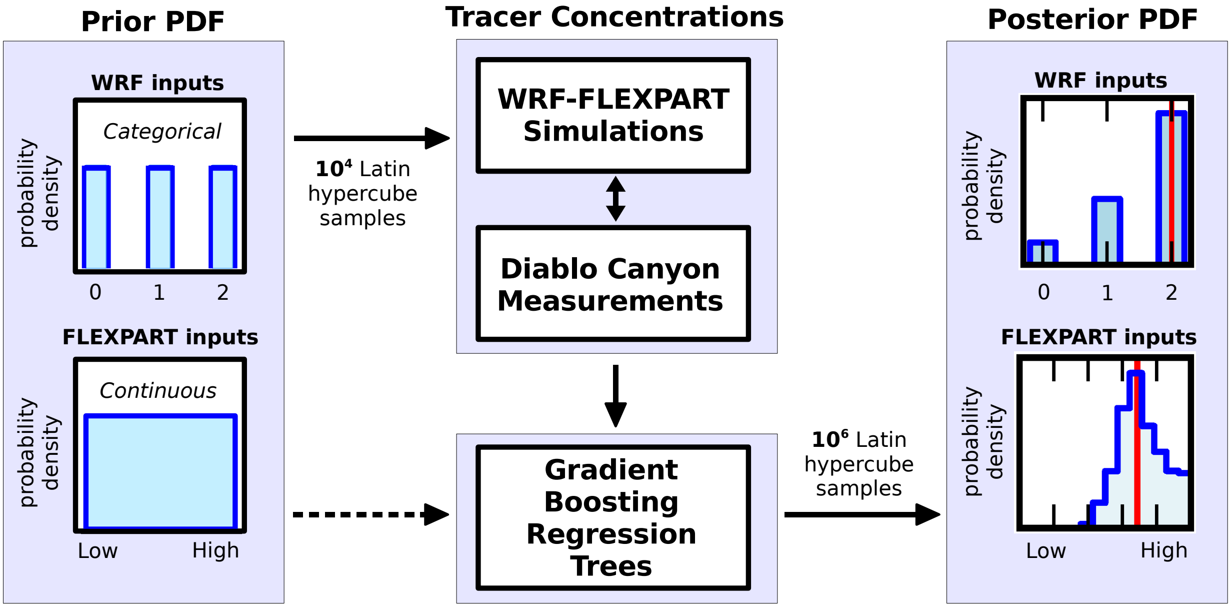
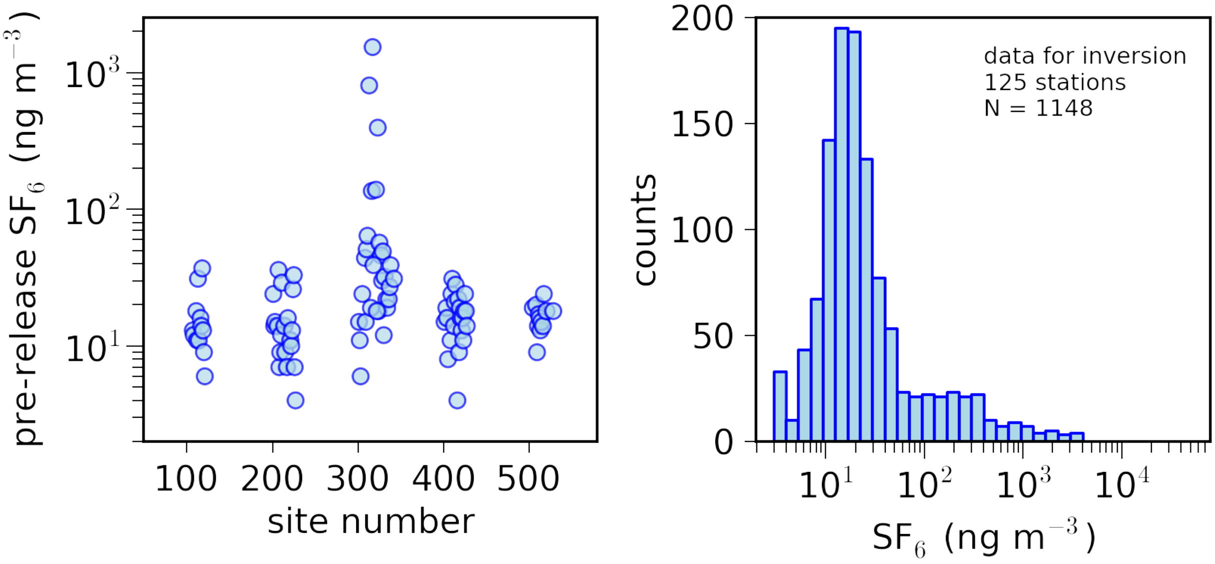


Figure 3: The upper panel shows the five nested domains used in WRF for simulating meteorological conditions for the Diablo Canyon nuclear power plant tracer release experiment (D1 to D5). The lower panel highlights the inner WRF domain (D5), the FLEXPART domain (dashed rectangle; longitudes of 120.954°W to 120.343°W, latitudes of 34.948°N to 35.389°N), the tracer release location (Diablo Canyon, red dot), and the measurement network (black dots) used for the tracer inversion. Four representative measurement sites (325, 330, 338, and 413) are highlighted (red dots). The colors on the maps denote the elevation above sea level. Different color scales are used for lower (0 to 1,000 m) and upper panels (0 to 3,000 m).



~~One-hour average SF_6 concentrations measured at the Diablo Canyon measurement sites during the tracer release test on the September 4, 1986. Pre-release concentrations (left hand side) show the effects of fugitive emissions at a subset of the site 300 locations (levels above 100 ng m^{-3}). Histogram of SF_6 concentrations from 1,148 measurements (right hand side) are used by the inversion algorithm to compute likelihood weights.~~

Figure 4: Bayesian inversion method for constraining the WRF and FLEXPART inputs. Samples are drawn from the uniform prior distribution on the left, and then evaluated in WRF-FLEXPART and compared to measurements. Gradient boosting regression trees are fit to model-measurement differences and used to infer the posterior distributions of the inputs.



~~Bayesian inversion method for constraining the WRF and FLEXPART inputs. Samples are drawn from the uniform prior distribution on the left, and then evaluated in WRF-FLEXPART and compared to measurements. Gradient boosting regression trees are fit to model-measurement differences and used to infer the posterior distributions of the inputs.~~

Figure 5: One-hour average SF$_6$ concentrations measured at the Diablo Canyon measurement sites during the tracer release test on the September 4, 1986. Pre-release concentrations (left hand side) show the effects of fugitive emissions at a subset of the site 300 locations (levels above 100 ng m^{-3}). Histogram of SF$_6$ concentrations from 1,148 measurements (right hand side) are used by the inversion algorithm to compute likelihood weights.

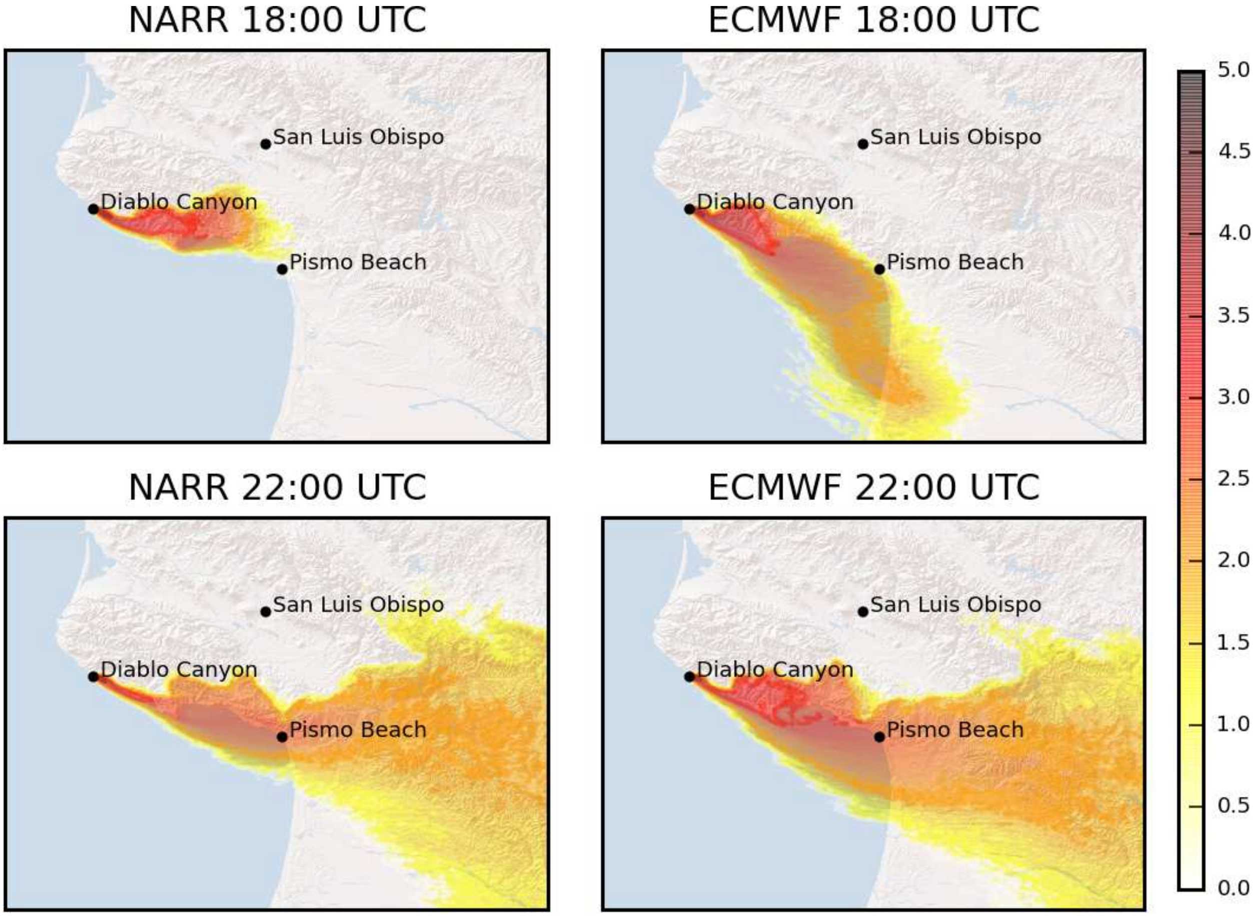


Figure 6: 30-minute average plumes of SF_6 simulated using FLEXPART with the actual release parameters (see Table 2). The plumes in the upper and lower panels are five and nine hours after the release, respectively, while those on the left and right use NARR and ECMWF reanalysis fields, respectively. The color scale shows the logarithm of the SF_6 concentrations between 10^0 ng m^{-3} and 10^5 ng m^{-3}.

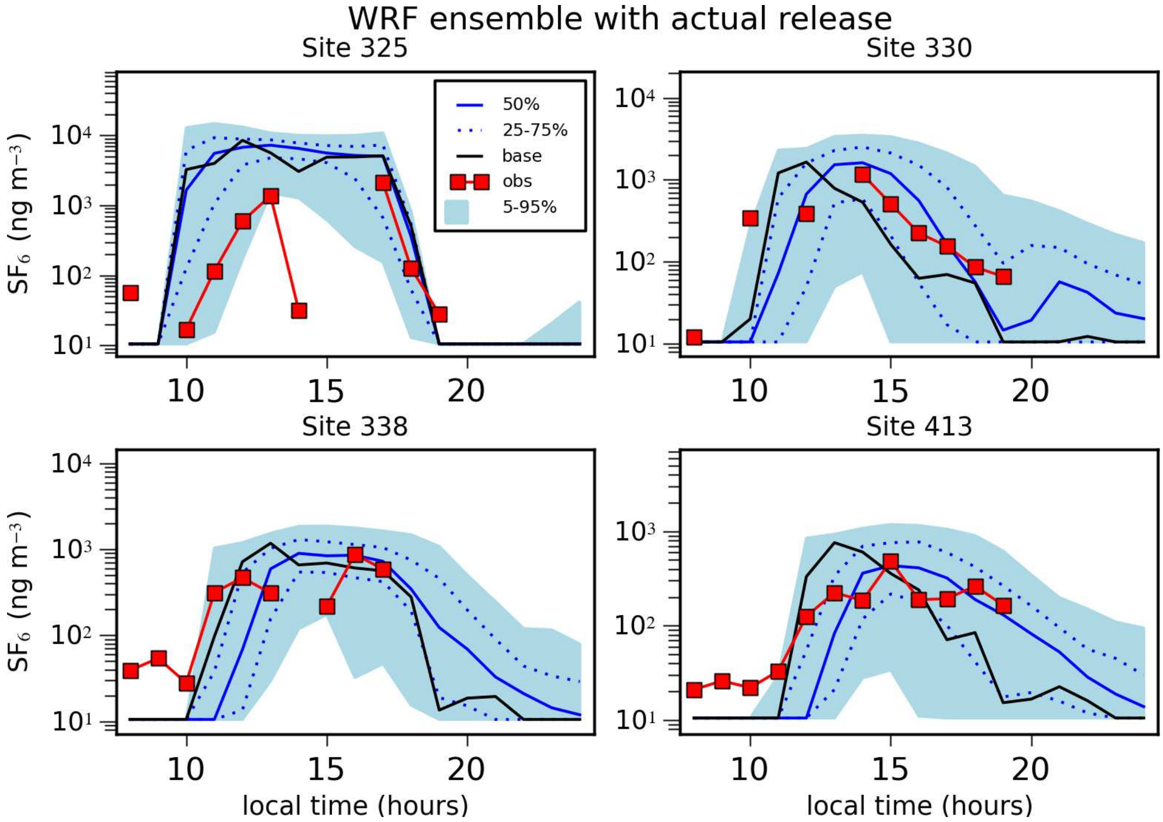


Figure 7: Time series of the distribution of SF$_6$ simulated using the actual Diablo Canyon release values and the 162 wind fields from the WRF ensemble. The time series are shown at four representative measurement locations (see Fig. 3). Different quantiles of the distribution are displayed (blue lines and area), as are the reference WRF-FLEXPART base case (black line), and the Diablo Canyon measurements (red squares). Local times correspond to the end of the one-hour average intervals. Note the different scales on the vertical axes.

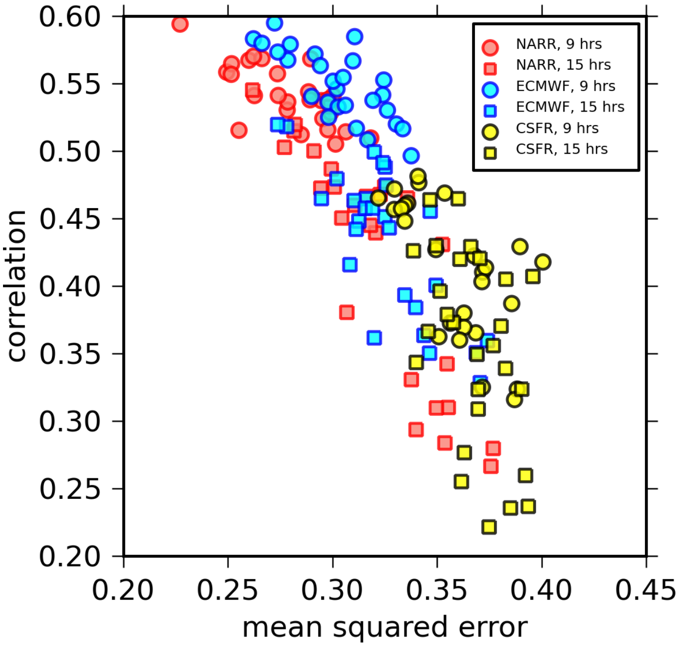


Figure 8: Correlation and mean-squared-error between the simulations and measurements of SF_6 for the cases using the actual Diablo Canyon release values and the 162 wind fields from the WRF ensemble. Colors and symbols are used to denote different reanalysis datasets and initialization times, respectively.

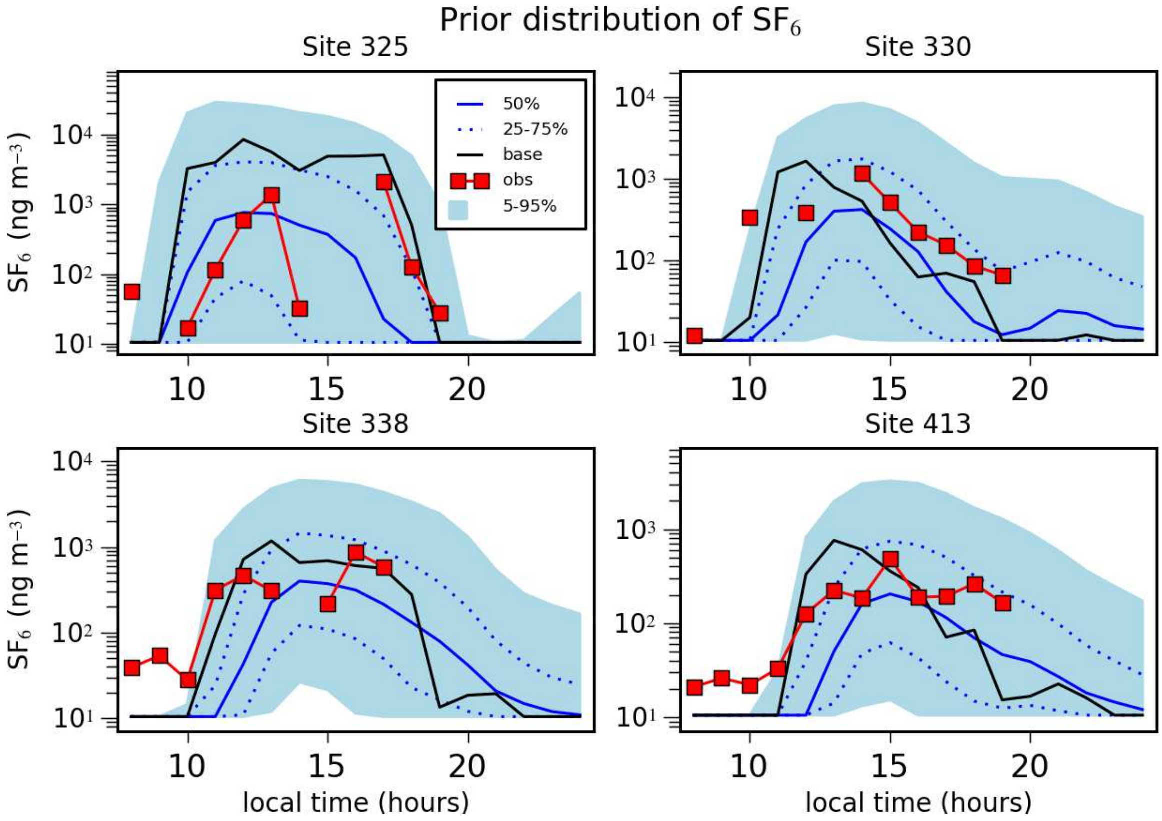


Figure 9: Time series of the prior probability distribution of SF_6 from the Latin hypercube ensemble of FLEXPART simulations at four representative measurement locations (see Fig. 3). Different quantiles of the probability distribution are displayed (blue lines and area), as are the base WRF-FLEXPART case using default and actual input values (black line), and the Diablo Canyon measurements (red squares). Local times correspond to the end of the one-hour average intervals. Note the different scales on the vertical axes.

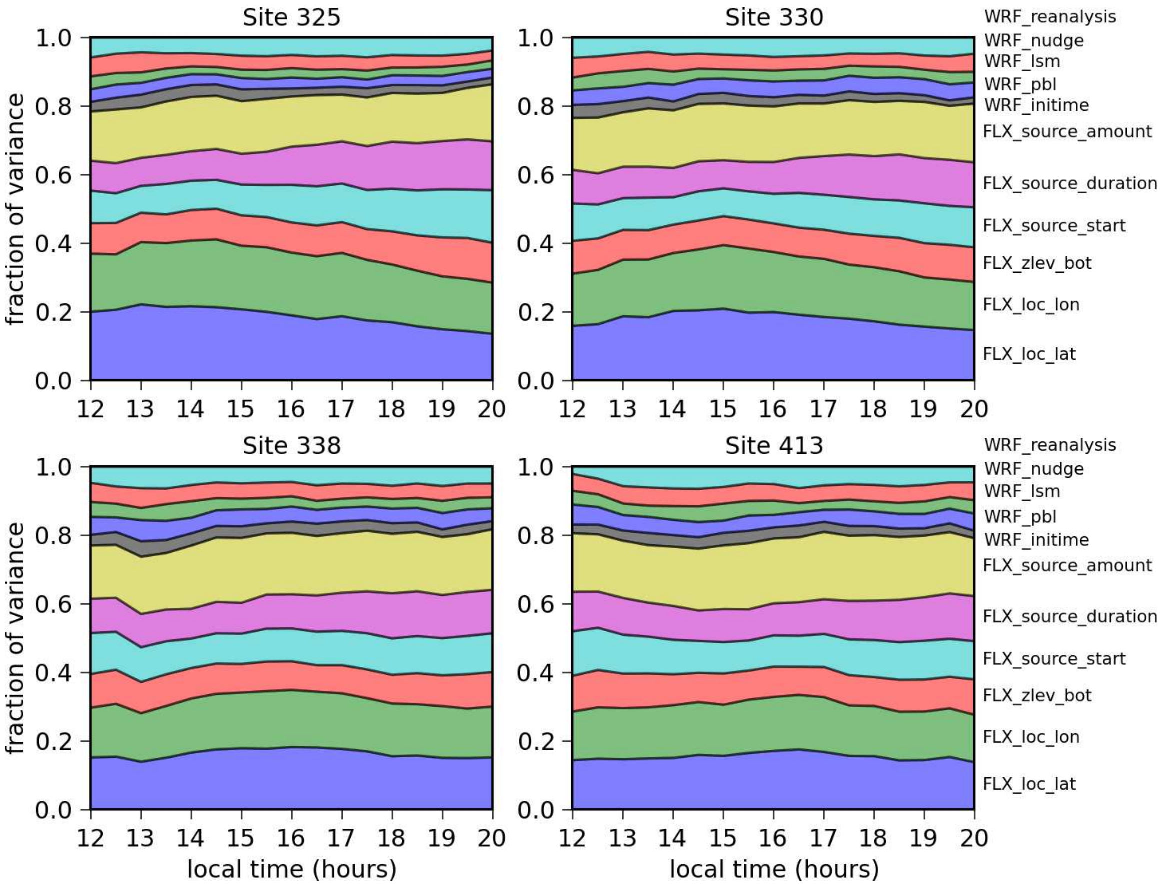


Figure 10: Time series of the SF_6 feature scores at the four representative measurement sites. Each colored band represents the fraction of the variance in the Latin hypercube ensemble (i.e., the prior distribution) caused by the parameters labeled on the right hand side. The WRF parameters are the upper five bands, while the FLEXPART parameters are the lower six bands.

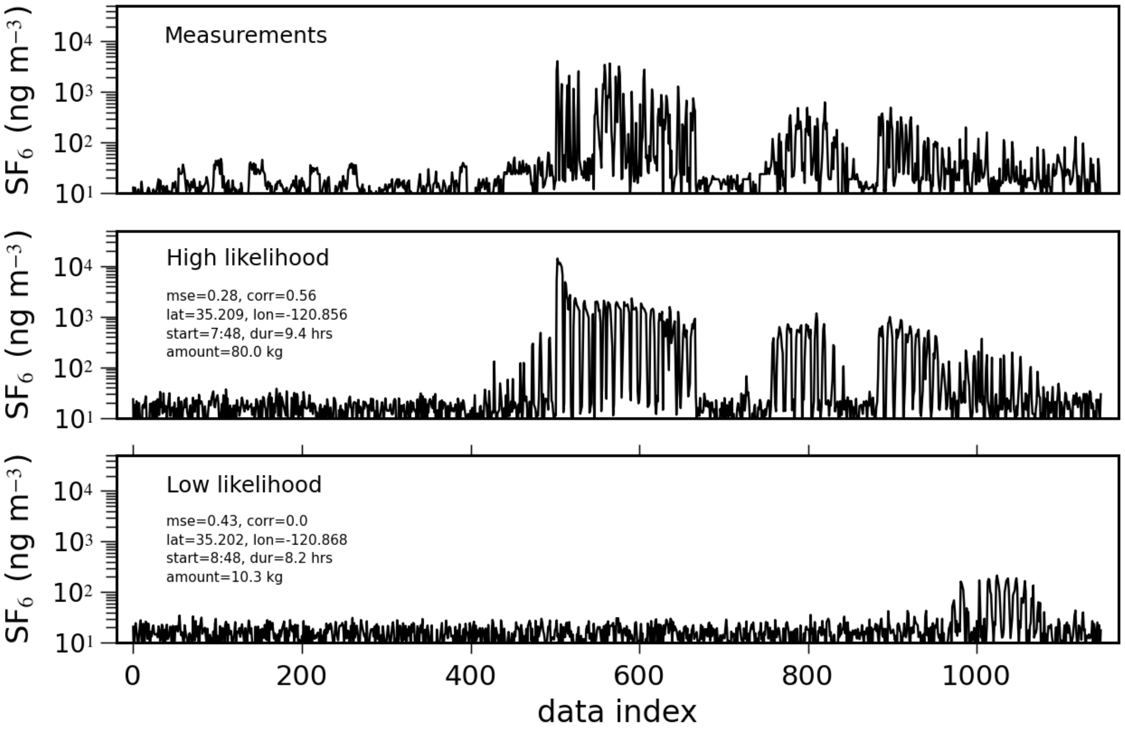


Figure 11: ~~Correlation and mean-squared-error between the simulations and measurements of~~ Diablo Canyon SF$_6$ ~~for~~ measurements (top row) are compared to representative high and low likelihood simulations from the ~~40,000 members of the WRF-FLEXPART~~ Latin hypercube ensemble (~~small blue dots~~middle and bottom rows, respectively). ~~The upper left portion~~ Data points are arranged by observation site number and time of ~~the figure shows the 50 best fitting simulations (red dots),~~ measurement. The simulation release parameters and ~~the estimated target~~ their *mse* and ~~covariance (1-to-3 standard deviation contours) in~~ *corr* values are listed on the ~~likelihood function~~left.

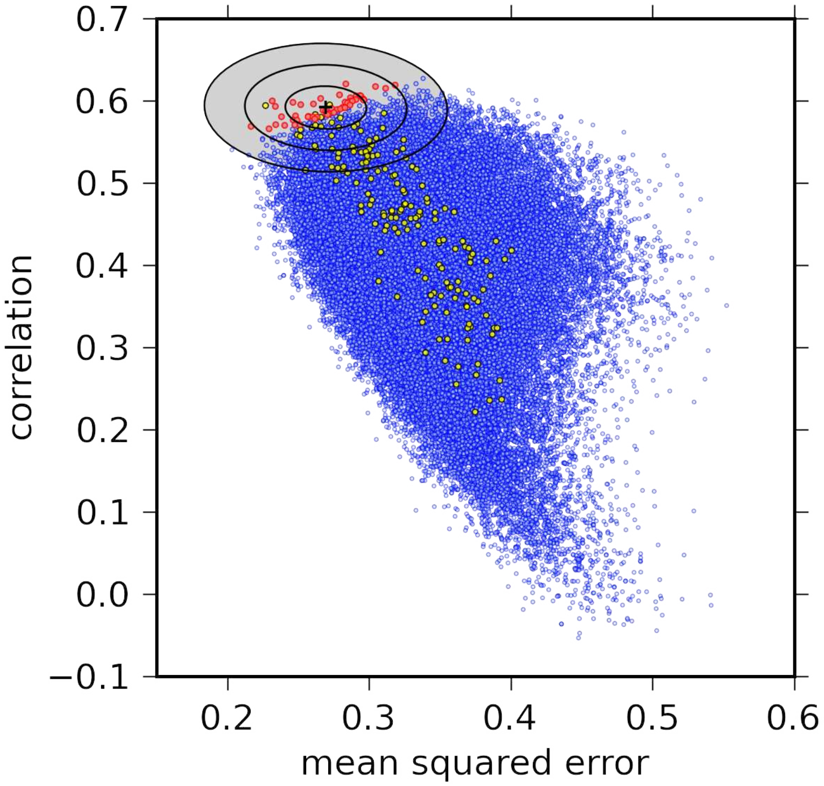


Figure 12: Correlation and mean-squared-error between the simulations and measurements of SF_6 for the 40,000 members of the WRF-FLEXPART ensemble (small blue dots). The upper left portion of the figure shows the 50 best fitting simulations (red dots), and the estimated target and covariance (1-to-3 standard deviation contours) in the likelihood function. The yellow dots show the *mse* and *corr* for the 162 reference simulations using actual release values (see Sect. 5.1 and Fig. 8).

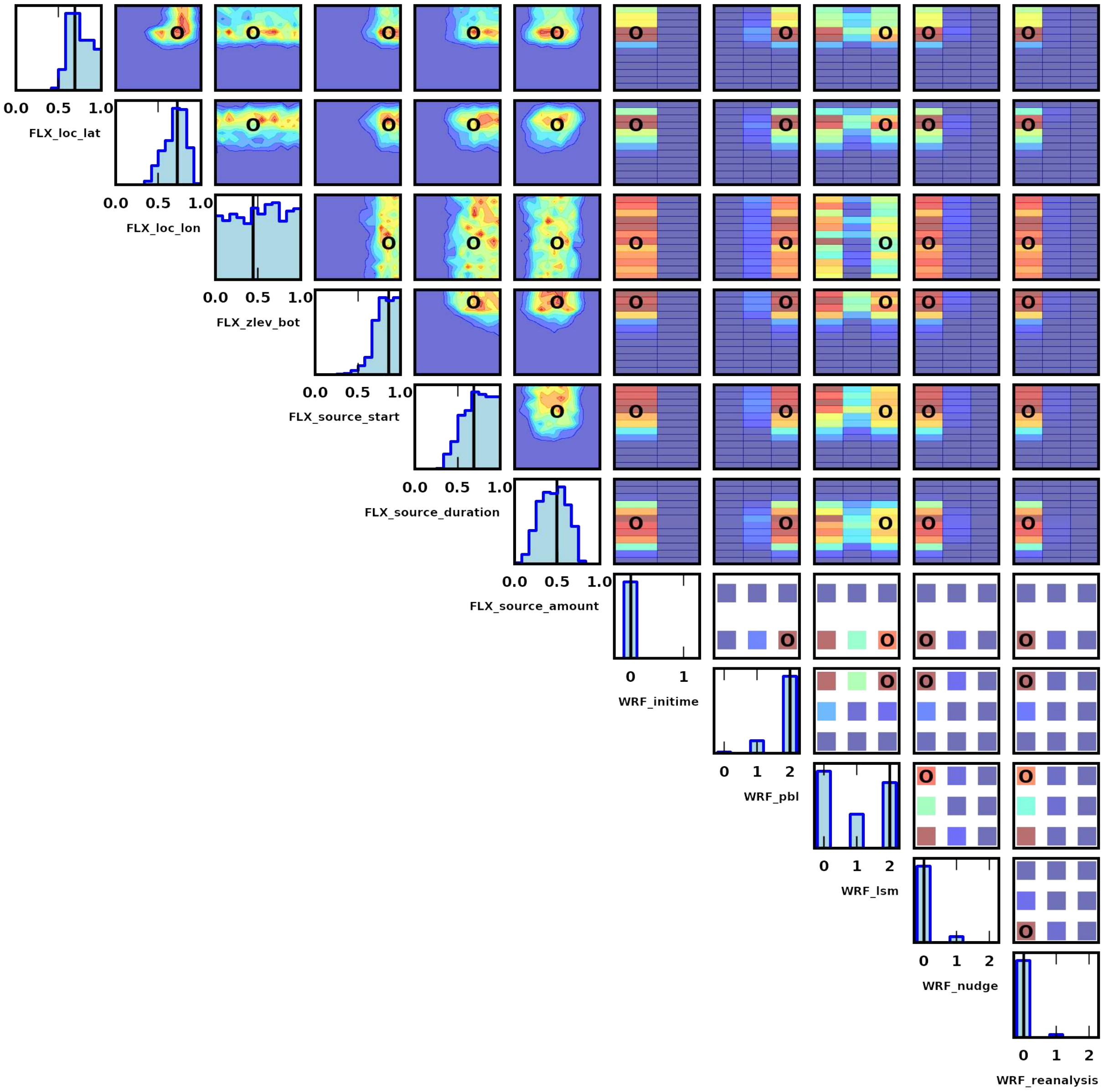


Figure 13: The marginal posterior distribution of WRF and FLEXPART parameters for the synthetic data inversion. Diagonal components show univariate continuous distributions for FLEXPART (top left) and univariate categorical distributions for WRF (bottom right). Off-diagonal components show bivariate distributions for the pair of parameters in the corresponding row and column. Probability density is normalized, with red colors denoting regions of high probability in the bivariate distributions. Known input values are shown by the black lines and circles in the diagonal and off-diagonal components, respectively. The FLEXPART parameter values have been scaled to [0,1] using the inversion ranges in Table 2 (log-scaling is used for the amount).

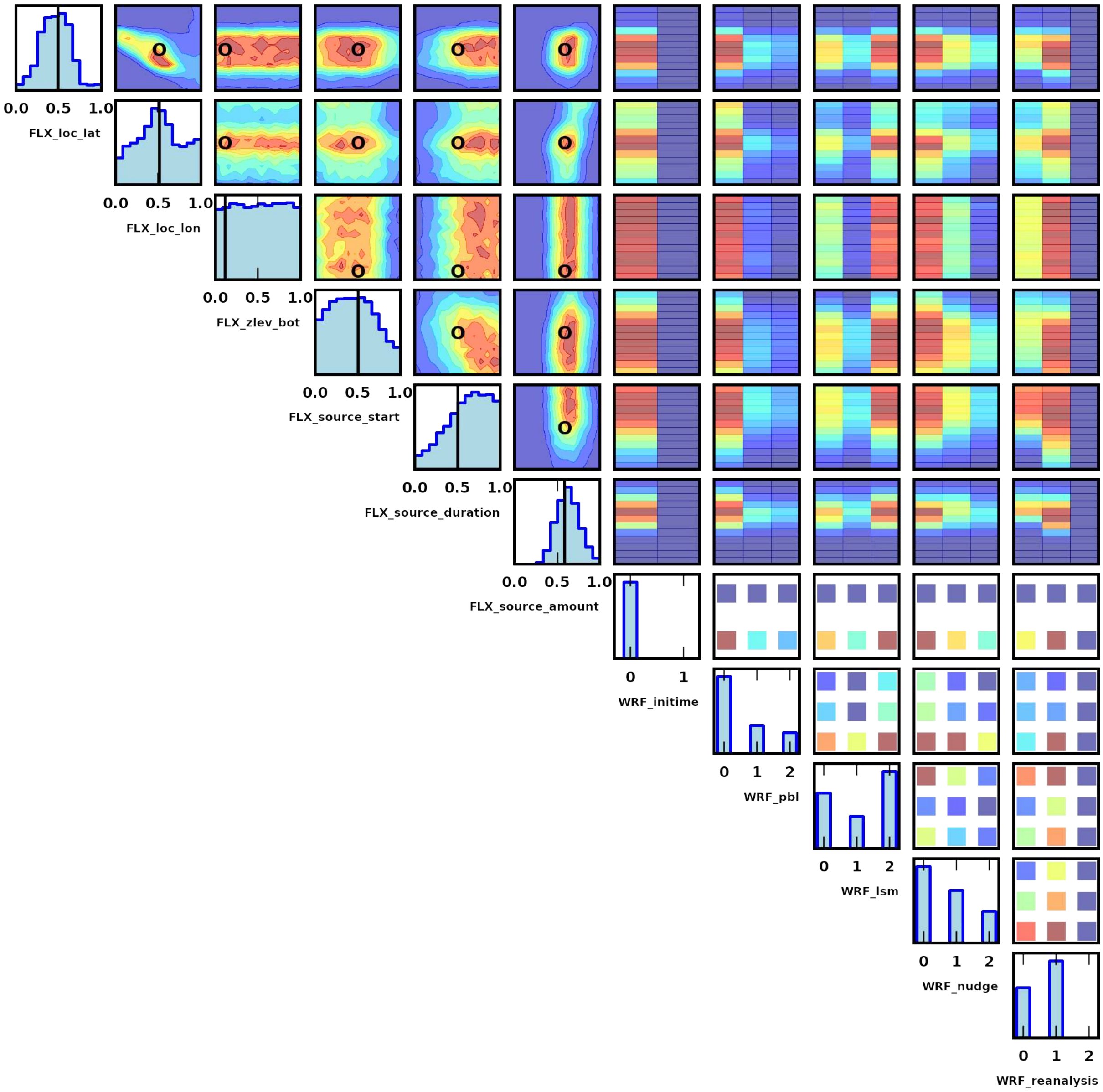


Figure 14: The marginal posterior distribution of WRF and FLEXPART parameters for the Diablo Canyon tracer data inversion. Actual source term input values used in the tracer release experiment are denoted by the black lines and circles. Refer to the caption in Fig. 13 for further information.

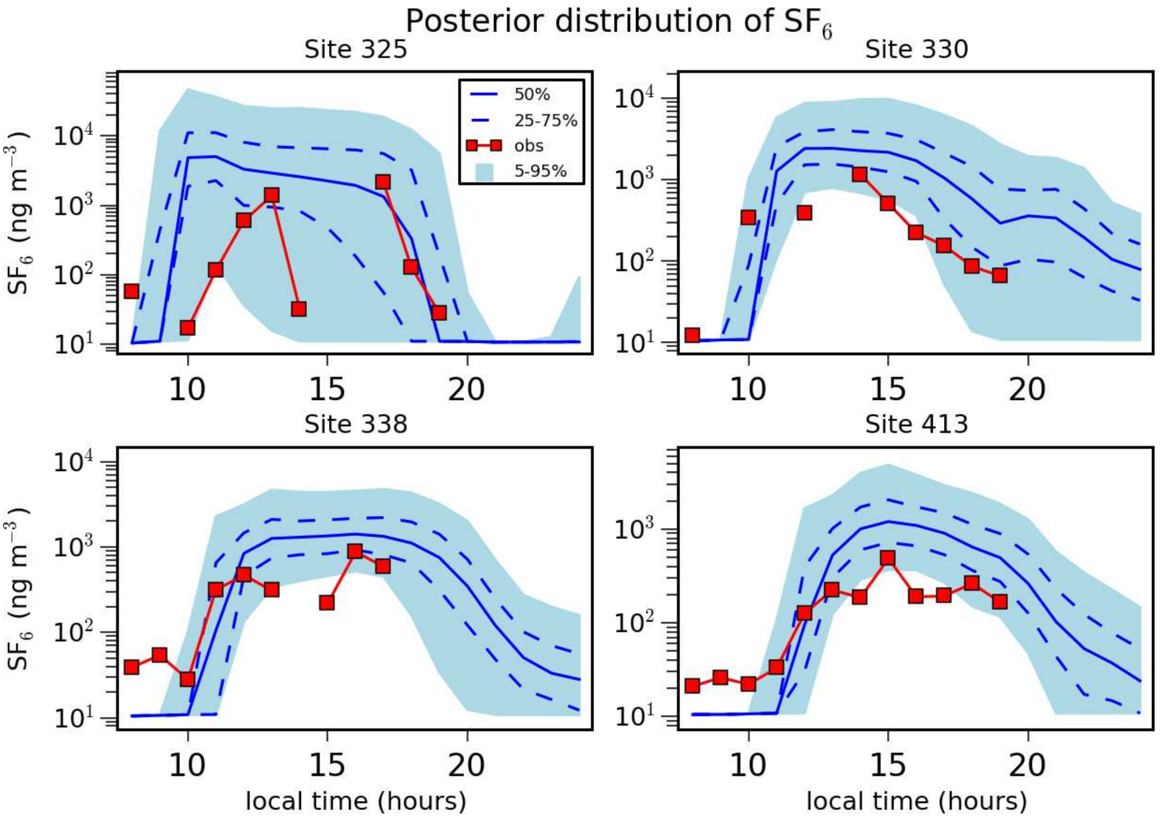


Figure 15: Time series of the posterior probability distribution of SF_6 at four representative measurement locations. Different quantiles of the probability distribution are displayed (blue lines and area), as are the tracer measurements (red squares).