NEW TOOLS FOR ATMOSPHERIC CHEMISTRY UTILIZING MACHINE LEARNING ON FIELD MEASUREMENTS

by

Mitchell P. Krawiec-Thayer

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: 01/18/2018

The dissertation is approved by the following members of the Final Oral Committee:
Frank N. Keutsch, Professor, Chemistry
Amir H. Assadi, Professor, Mathematics
Timothy H. Bertram, Associate Professor, Chemistry
John Wright, Professor, Chemistry
J.R. Schmidt, Associate Professor, Chemistry

© Copyright by Mitchell P. Krawiec-Thayer 2018
All Rights Reserved

This work is dedicated to the continuing process of scientific introspection...

ACKNOWLEDGMENTS

There are a great number of communities that have supported me over the course of this journey. First and foremost, none of this would have been possible without the consistent encouragement and help of my family. Endless thanks to my partner Allison, whose unconditional love and optimism brought light to many rough times. My life over the past 5 years has been immensely enriched by the communities at Christ Presbyterian Church and Avalon Coöperative House, who have been my second families in Madison.

I thank Jennifer B. Kaiser and Kate M. Skog for years of patient mentoring. Jen, Kate, and Matthew R. Dorris were absolutely instrumental in the collection these data and surviving long field campaigns. I also appreciate the companionship of Sean Staudt, who accompanied me through academic transitions and many Friday swap runs.

Many thanks to Frank N. Keutsch for years of adventures and allowing me the freedom to explore new directions and ideas. In recent years, Timothy H. Bertram welcomed me into his group and provided much valuable input and feedback.

Last and certainly not least, I am immensely grateful to Amir H. Assadi for his generous kindness and support. Without his insight and guidance, many of the ideas in this thesis would never have approached fruition. His enthusiastic desire to help every student achieve their dreams has launched uncountable interdisciplinary projects and epitomizes the ideals of academic collaboration.

NEW TOOLS FOR ATMOSPHERIC CHEMISTRY UTILIZING MACHINE LEARNING ON FIELD MEASUREMENTS

Mitchell P. Krawiec-Thayer

Under the supervision of Professor Frank N. Keutsch
At the University of Wisconsin-Madison

Atmospheric chemistry and meteorological measurements produce large heterogeneous datasets that capture complex physical phenomena. Many of the models and analyses carried out on these data fundamentally consist of pattern recognition, regression, and classification tasks. Such activities are extremely amenable to improvement and/or automation with machine learning. My thesis details new machine learning-based tools that I developed during the analysis of measurements collected by the Keutsch group during our field campaigns in Finland, Brazil, and the western United States.

Large collaborative datasets inevitably include some times during which not all instruments' measurements are available (due to calibration/zeroing periods, maintenance, etc), and these gaps must be addressed prior to any model or analysis that requires continuous inputs. I discuss the development of several multivariate imputation methods that fill gaps in one data source based on information learned from the other measurements recorded simultaneously. This approach is demonstrated on both ground and flight data using techniques such as lazy learners, regression learners, and artificial neural networks.

The concentrations of chemical pollutants near the ground depend on the dynamic height of the lowest layer of the atmosphere. My thesis describes a new method for robust identification of atmospheric structure through novel application of cluster evaluation measures. Finally, I combine this structural information with the chemical measurements to emulate spatial variability in retrievals from satellite instruments.

CONTENTS

Contents iv

List of Tables vii

List of Figures viii

1 Introduction 1

1.1 Imputation of missing measurements 2

1.2 Cluster evalutaion to determine atmospheric structure 3

1.3 Data visualization from satellite perspective 4

2 Machine Learning Methods for Imputing Missing Atmospheric Measurements in Ground & Flight Data 6

2.1 Introduction 7

2.2 Methods 8

2.2.1 Univariate Methods 8

2.2.1.1 Measurement Mean 8

2.2.1.2 Interpolation 9

2.2.1.3 Diel Average 9

2.2.2 Multivariate Methods 9

2.2.2.1 Preliminary processing 9

2.2.2.2 Adaptive Average 10

2.2.2.3 Regression Learning 11

2.2.2.4 Artificial Neural Networks 11

2.2.3 Data Sets 12

2.2.3.1 Ground Measurements from GoAmazon 12

2.2.3.2 Flight Measurements from SONGNEX 12

2.2.4 Performance testing 12

2.3 Results 13

2.3.1 Feature Selection 13
2.3.2 Neighbor selection 13
2.3.2.1 GoAmazon examples (ground) 13
2.3.2.2 SONGNEX examples (flight) 15
2.3.3 Number of Training Points 16
2.3.4 Adaptive Average Weighting 20
2.3.5 Neural Network Architecture 21
2.3.6 Neural Network Replicates 21
2.3.7 Intercomparison 26
2.4 Conclusions 29
2.5 Future work 33
2.6 Acknowledgements 34

3 Identification of Planetary Boundary Layer Height from Self-Similarity in Vertically-Resolved Measurements 35
3.1 Introduction 36
3.2 Theory 38
3.2.1 Data Preparation 38
3.2.2 Distances 38
3.2.3 Forward clustering algorithms 39
3.2.4 Cluster Evaluation Methods 40
3.2.4.1 Silhouette criterion values 41
3.2.4.2 Calinski-Harabasz criterion values 42
3.3 Results 42
3.3.1 Forward Clustering 42
3.3.2 Comparison of Indices 43
3.4 Conclusions 45
3.5 Future Work 48

4 Assessment of Satellite Capabilities for Discerning HCHO, O_3, and NO_2 Enhancements from Multiple Sources 49

4.1 Introduction 50
4.2 Methods 51
4.2.1 Measurements 51
4.2.2 Analysis 51
4.2.2.1 Converting mixing ratio to column density 51
4.2.2.2 Satellite Emulation 52
4.3 Results and Discussion 54
4.3.1 Flight and Observations 54
4.3.2 Spatial Resolution 58
4.3.3 Chemical Precision 62
4.4 Satellite Comparisons 63
4.4.1 OMI 65
4.4.2 TROPOMI 65
4.4.3 TEMPO 65
4.5 Conclusions 66
4.6 Future Work 67
4.7 Acknowledgements 67

5 Future Work 68
5.1 Support vector machines 68
5.2 Super Resolution 68
5.3 Markov chain methods & fuzzy comprehensive decision model 69
5.4 Note on cloud computation 70

Bibliography 71

LIST OF TABLES

4.1 Emulations of satellite instruments are based off of TROPOMI specifications and TEMPO expected performance. 54
4.2 Precision necessary to a) discern broad plumes and b) resolve finer details 63

LIST OF FIGURES

2.1 Data points used for estimation of 'missing' HCHO on 19-Sept 2014 (*a* & *b*) and 26-Sept 2014 (*c* & *d*) during the GoAmazon campaign. The top panel for each day shows a masked point to be imputed (red marker), the data points that are used for the diel average (yellow markers, at the same time each day), and the times with similar conditions identified for the adaptive average (blue markers, from nearest-neighbor search). The top panel for each day shows the timeseries over the course of the dry season, while the bottom panel for each shows the same data plotted is a function of time of day. 14

2.2 Data points used for estimation of a 'missing' HCHO measurement in the free troposphere during the SONGNEX research flight. The red marker indicates the masked point to be imputed. The cyan markers show the points that would be used for imputation by linear interpolation, and the blue markers show points with similar conditions identified for the adaptive average. 15

2.3 Data points used for estimation of a 'missing' HCHO measurement from the planetary boundary layer during the SONGNEX research flight. The red marker indicates the masked point to be imputed. The cyan markers show the points that would be used for imputation by linear interpolation, and the blue markers show points with similar conditions identified for the adaptive average. The background shows the (smoothed) spatial distribution of ozone in the planetary boundary layer, which was the strongest predictor for HCHO. 17

2.4 Plots showing error analysis used for tuning the number of points to be included in the adaptive average. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set. 18
2.5 Plots showing error analysis used for tuning the number of points to train the neural network. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set. 19
2.6 Plots showing error analysis used for tuning the number of points to train the regression learner. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set. 20
2.7 Distance values of nearest neighbors as a function of their distance from the masked point (calculated from the z-scores as in equation 2.5. The cumulative mean is shown with the stepwise addition of each point. . 21
2.8 Plots showing error analysis used for tuning the number hyperparameter ρ, the distance exponenent in the weighting function. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set. 22

2.9 Plots showing error analysis used for testing various NN configurations on the GoAmazon data set (left) and SONGNEX data set (right). The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set. 23
2.10 Diagram of the NN configuration generated for imputing HCHO in the GoAmazon data set. 24
2.11 Statistical analysis of NN outputs for 100 training replicates for the same point. The histogram on the left shows the distribution of estimations. The upper right figure shows the value of each estimate and stepwie calculation of the mean (blue trace). The bottom right figure shows the change in mean with each successive addition (blue trace), along with the threshold for stable mean (dotted black line) and the replicate at which the criteria are met (solid black line). 25
2.12 Number of NN replicates required to meet halting conditions (<0.5% change in mean with the inclusion of 6 successive replicates) in the GoAmazon data set. The 65% that halted at the minimum number of estimates (20) are not shown; these are statistics for the 35% that required more replicates. 26
2.13 Pointwise intercomparison of imputation methods with GoAmazon ground measurements on the top and SONGNEX flight measurements on the bottom. A data point whose reconstructed value matches the true masked value falls on the 1:1 diagonal. The further from the diagonal, the more that a point was under/over predicted (to the right and to the left, respectively). 27

2.14 Intercomparison of imputation methods for HCHO field data (GoAmazon left, SONGNEX right). The top panels show mean absolute error (MAE) at varying concentrations, with with the pointwise average MAE for each method in the legend text. The middle panels show the bias as a function of concentration, with the average bias in the legend text. The bottom panels show the distribution of HCHO in each data set, for reference. 28
2.15 Intercomparison of adaptive average and diel average imputation estimates for isoprene during GoAmazon. A data point whose reconstructed value matches the true masked value falls on the 1:1 diagonal. The further from the diagonal, the more that a point was under/over predicted (to the right and to the left, respectively). 30
2.16 Intercomparison of imputation methods for ozone field data (GoAmazon left, SONGNEX right). The top panels show mean absolute error (MAE) at varying concentrations, with with the pointwise average MAE for each method in the legend text. The middle panels show the bias as a function of concentration, with the average bias in the legend text. The bottom panels show the distribution of ozone in each data set, for reference. 31
2.17 Intercomparison of imputation methods for fine particulate matter (optical diameter 0.004-1.0 μm). The top panels show mean absolute error (MAE) at varying concentrations, with with the pointwise average MAE for each method in the legend text. The middle panels show the bias as a function of concentration, with the average bias in the legend text. The bottom panels show the distribution of ozone in each data set, for reference. 32

3.1 Measurements from the first SONGNEX vertical profile. Specific humidity (SH), potential temperature (θ), ozone (O_3), and formaldehyde (HCHO) exhibit discontinuities near $z' \sim 650$ m. 36

3.2 Measurements from the first SONGNEX vertical profile. Specific humidity (SH), potential temperature (θ), and ozone (O_3). The data are colored by height to highlight the separation between clusters near $z' \sim 650$ m. 37
3.3 L2-norm similarity matrix for a PBL→FT ascent. Clusters appear as self-similar blocks along the diagonal, so z' is visually apparent. 39
3.4 Specific humidity (SH), potential temperature (θ), and ozone (O_3) for the first ascent on the SONGNEX 2015.04.14 flight. The right pane shows orange $\chi_\alpha(h)$ and yellow $\chi_\beta(h)$ traces, along with red line indicating h_{est} 43
3.5 Specific humidity (SH) and potential temperature (θ) for the first three ascents on the SONGNEX flight. The silhouette values and the Calinski-Harabasz indices for each partitioning are shown at left, along with the height corresponding to their global maximum. 44
3.6 The Calinski-Harabasz index method is used to find z' for the first SONGNEX ascent using three mutually-exclusive subsets of measurement types: meteorological (top), trace gases (middle), volatile organic compounds (bottom). 46
3.7 Relative humidity (RH), potential temperature, and hydroxyl reactivity measurements (kOH) during the first three ascents on the PEGASOS flight are evaluated by the Calinski-Harabasz value to find z'. 47

4.1 The height of the PBL during a given profile is easily identified from discontinuities in specific humidity (SH) and potential temperature (θ). These measurements from the first PBL→FT ascent are shown here as a function of height above ground level, alongside HCHO for comparison. 53
4.2 Top-down view of the 2015.04.13 flight track, cropped to the portion of the raster pattern that will be used for satellite emulation. The flight track is colored by formaldehyde, and the three regions of interest are indicated. 55
4.3 Histograms of ambient concentrations in the biomass burning (blue), urban outflow (red), and background (gray) regions, sorted by chemical starting clockwise from upper left: HCHO, O_3, CH_4, NO_2. 56

4.4 HCHO is a tracer for particulate matter in the biomass burning plume 56
4.5 A 3-dimensional view of the flight track, showing the spirals at each corner. high formaldehyde concentrations and variability are limited to the planetary boundary layer. 57
4.6 Vertical distribution of HCHO over the entire flight. High concentrations and spatial variability are seen in the PBL, while the FT has lower concentrations and is relatively homogenous throughout the flight. . . 58
4.7 High-resolution (1x1 km) distribution of HCHO in the PBL in the region of interest. This shows the emulated satellite perspective with no imposed constraints limiting spatial resolution or chemical precision. . 59
4.8 The effect of spatial resolution on the ability to discern chemical enhancements, shown for HCHO binned to 4 pixel footprints (clockwise starting from upper left: 20x20km, 10x10km, 5x5km, 2x2km). The color scale is the same in each figure, and the arrows on the colorbar indicate attenuation of the observed concentration ranges as variations are averaged into larger pixels, losing information about the extremes. . . . 60
4.9 Tropospheric O_3 at 5x5 km resolution 61
4.10 Tropospheric NO_2 at 5x5 km resolution 61
4.11 The effect of chemical precision (signal uncertainty) on the ability to discern chemical enhancements, shown for HCHO binned from most coarse to most precise (clockwise starting from upper left: $2x10^{16}cm^{-2}$, $8x10^{15}cm^{-2}$, $4x10^{15}cm^{-2}$, $2x10^{15}cm^{-2}$) 62
4.12 Expected TROPOMI (left) and TEMPO (right) perspectives on the region of regard, emulating limitations described in Table 4.1. From top to bottom, HCHO, O_3, and NO_2. The arrows on the colorbars indicate attenuation of the observed concentration ranges as variations are averaged into larger pixels. 64

1 INTRODUCTION

Atmospheric chemistry and meteorological measurements produce large heterogeneous datasets that capture complex physical phenomena. There is a continuous influx of global data from remote-sensing satellites, complemented by in-situ measurements such as height-resolved profiles from radiosondes (weather balloons) and observations from sensors at the ground level. These diverse data are collected at a staggering rate, on the order of terabytes per hour — much faster than it can be analyzed and studied.

Many of the models and analyses carried out on these data fundamentally consist of pattern recognition, regression, and classification tasks. Such activities are extremely amenable to improvement and/or automation with machine learning. They adapt well to the challenges that arise when analyzing and learning from noisy, incomplete, and high-dimensional field data.

Chapter 2 of this thesis details new tools for filling gaps in atmospheric measurements (adaptive averaging, neural networks, regression learning) through machine learning on other timepoints sampling similar conditions. This innovation improves on earlier studies on imputing geostationary measurements [27] by training each learner only the datapoints containing the most relevant information (e.g. a neural network learning to estimate data points on hot dry afternoons does not benefit from training on cold rainy midnight data) since indiscriminate learning to minimize prediction error for *all* conditions inadvertently increases prediction learning error for the desired conditions. This work also represents the first study on multivariate imputation methods for flight data.

Chapter 3 details a novel application of cluster evaluation measures for identifying atmospheric structure from in-situ measurements. Current methods for identifying the boundary layer height from in-situ data rely on identifying particular features in the profiles such as: gradients in specific humidity (SH), potential temperature (θ), or derived Richardson number.[56][53] These focus on single characteristics, their performance can be

resolution dependent (especially humidity-based methods), and variations in structure can lead to false assignments. This work details the first application of methods from cluster analysis toward identifying atmospheric structure. By virtue of learning on the meteorological measurements and other quantities impacted by the boundary layer height, these analyses incorporate the information from multiple traditional techniques and agument this with additional information. Consequently, they are significantly more robust to anomolies in atmospheric structure, insensitive to measurement resolution, and perform well on any set of features that exhibits decoupling between the boundary layer and the free troposphere.

The results of chapters 2 and 3 are combined in chapter 4 to visualize flight data in a novel manner to emulate the spatial variability in satellite instrument retrievals and study the information present at varying scales. The spatial resolution and instrument precision of chemical measurements impact which atmospheric phenomena are discernible from remote sensors,[20] and this work demonstrates a method for studying these effects directly from flight data, without requiring an intermediate model. The analysis is applied to formaldehyde, ozone, and nitrogen dioxide data to assess 'minimum requirements' for discerning various phenomena. Lastly, the specifications and expected performance of TROPOMI (2017 launch) and TEMPO (upcoming) are incorporated to emulate the instruments and anticipate which phenomena they will be capable of detecting.

1.1 Imputation of missing measurements

Large collaborative datasets inevitably include some times during which not all instruments' measurements are available (due to calibration/zeroing periods, maintenance, etc), and these gaps must be addressed prior to any model or analysis that requires continuous inputs. Common univariate approaches for replacing missing data include linear interpolation or substitution of a diel average, depending on the duration and nature of the gap.

Chapter 2 details three multivariate imputation techniques that apply supervised machine learning techniques to estimate a missing data point based on information contained in simultaneous collocated chemical and meteorological measurements. The data points that do not contain gaps provide the training set for methods that estimate the target feature (measurement type with missing data). Stepwise feature selection discards irrelevant and redundant attributes, and identifies a subset of variables with the most predictive value for a given measurement type in the subject data set.

An adaptive tracer method efficiently estimates the missing feature from a weighted average of nearest neighbors in the predictor coordinate space. Binary tree regression on nearby data points provide another option for estimating the missing value. Lastly, a feedforward neural network is trained by backpropogation to specialize in prediction of the target feature.

Model performance is quantified by leave-one-out cross validation on two data sets. The GoAmazon 2014/15 dry season ground measurements [36] are used to assess performance with geostationary timeseries. A research flight during the Shale Oil and Natural Gas Nexus (SONGNEX 2015, WP-3D Orion) field campaign [42] provides a test set for application to data with temporal and spatial variation. In both cases, the adaptive multivariate methods yielded better estimates than the univariate approaches.

1.2 Cluster evalutaion to determine atmospheric structure

The height of the planetary boundary layer (PBL) significantly impacts pollutant concentrations near the ground by affecting the extent of vertical mixing. The structure of the atmosphere changes over the course of the day, and its behavior must be measured or estimated for modeling or forecasting. Because the PBL and free troposphere (FT) are largely decoupled for many physical/chemical quantities, variation between layers is much greater than variation within a layer.

Chapter 3 describes how techniques from cluster analysis can be used to identify the height of the PBL from vertically-resolved measurements, such as those obtained by radiosondes or research flight ascents/descents. This is accomplished by identifying the global maximum in cluster evaluation indices calculated for hypothetical partitions at each height measured during a given profile. Evaluation using silhouette values [29][47] and Calinski-Harabasz indices [8] both yield clear maxima in agreement. Techniques and shortcomings of forward clustering algorithms (centroid-based, density-based, and Gaussian mixtures) are discussed.

This approach is demonstrated on flight data sets from the Pan-European Gas Aerosol Climate Interaction Study (PEGASOS 2013, zeppelin) and SONGNEX field campaigns. This analysis requires no parameterization and is shown to be robustly applicable to multiple classes of measurements (metorological, trace gases, volatile organic compounds, reactivity).

1.3 Data visualization from satellite perspective

Instruments to be deployed on new and upcoming satellites, such as TROPOMI (2017 launch) and TEMPO (2018/19 projected launch), will be capable of measuring air quality-relevant species at unprecedented spatial and temporal scales,[4][7][54][66] however studies on the degree of detail that this will afford are lacking. Chapter 4 addresses the question: Given a region with a complex scene featuring signatures from multiple types of anthropogenic activity, what spatial resolution and chemical precision are necessary to resolve plumes and distinguish between emissions types? These specifications are studied for formaldehyde (HCHO), nitrogen dioxide (NO_2), and tropospheric O_3.

Chemical and physical parameters measured via aircraft in the boundary layer and free troposphere (FT) during the Shale Oil and Natural Gas Nexus (SONGNEX 2015) field campaign are employed to view chemical enhancements from biomass burning and urban

outflow over the region northeast of Denver, Colorado. The spatially and temporally resolved in-situ data are used to calculate the planetary boundary layer contributions to the column densities for HCHO, O_3, and NO_2. The converted data are mapped with varying constraints imposed to mimic equipment limitations. Pixel footprint resolution is probed using 2D spatial bins, and the loss of detail due to uncertainty in measurements is emulated by 1D signal bins to mask gradients that are finer than the instrumental precision.

First, the spatial resolution and chemical precision limits are studied for each species. Second, the scene is emulated using the specifications for TROPOMI [54][18][57] and expected performance for TEMPO, [66] in order to assess the degree to which their retrievals will be able to discern the signatures of various activities, and to ascertain the information that may be derived from trace gas enhancements.

2 MACHINE LEARNING METHODS FOR IMPUTING MISSING ATMOSPHERIC MEASUREMENTS IN GROUND & FLIGHT DATA

ABSTRACT: Atmospheric datasets typically include some data points where not all measurements are available. Interruptions in data collection occur for variety of reasons such as calibration/zeroing periods, loss of power, maintenance, etc. Common univariate approaches for replacing missing data include linear interpolation or substitution of a diel average, depending on the duration and nature of the gap.

This intercomparison assesses three multivariate imputation techniques that apply supervised machine learning techniques to estimate a missing data point based on information contained in simultaneous collocated chemical and meteorological measurements. The data points that do not contain gaps provide the training set for methods that estimate the target feature (measurement type with missing data). Stepwise feature selection discards irrelevant and redundant attributes, and identifies a subset of variables with the most predictive value for a given measurement type in the subject data set.

An adaptive tracer method efficiently estimates the missing feature from a weighted average of nearest neighbors in the predictor coordinate space. Binary tree regression on nearby data points provide another option for estimating the missing value. Lastly, a feedforward neural network is trained by backpropogation to specialize in prediction of the target feature.

Model performance is quantified by leave-one-out cross validation on two data sets. The GoAmazon 2014/15 dry season ground measurements are used to assess performance with geostationary timeseries. A research flight during the SONGNEX 2015 field capaign provides a test set for application to data with temporal and spatial variation. In both cases, the adaptive multivariate methods yielded better estimates than the univariate approaches.

2.1 Introduction

Atmospheric field campaigns and measurement stations produce large sets of collocated time series data capturing a variety of chemical and meteorological features. Missing data is a common occurrence for a variety of reasons. The gaps in data vary in duration and are usally not correlated between different instruments (e.g. the ozone measurements continue while the formaldehyde instrument is calibrating).[27]

Models frequently require continuous inputs, and analysis quality can depend on data set completeness. Gaps in a measurement must be addressed by either imputing estimates for the missing data or removing the timepoint(s) from analysis.[14] In datasets with many measurements, the amount of time that all instruments are simultaneously producing data is typically very limited. Consequently, good methods for filling in data are critical for practical application.

Several univariate methods exist for approximating missing values. The most straightforward and least accurate method is to substitue the mean value for a given feature in place of missing measuremens. For short gaps, data is often be imputed using linear interpolation. For gaps in ground data, common practice is to replace missing points with the 'diel average' calculated from the mean of measurements collected at the same time on other days.

While convenient, these approaches produces a rough first approximation and discard the valuable information contained in the measurements of other attributes. The multivariate methods presented here use the simultaneous non-missing measurements to identify relevant similar points and guide reconstruction of the missing data based on relationships contained within a given set.

A general procedure for implementation of machine learning methods is described below, independent of any particular data set or measurement type. The advantage of using a framework that applies variable selection and learning theory to each data replacement task is that the resulting models are highly specialized for estimating the target feature

based on the relationships captured in the available measurements.

The intercomparison figures include results from all of the above methods to guide selection of the proper tool(s) for a particular imputation task. Whether applying the standard univariate imputation techniques or using the machine learning methods described below, it is always the responsibility of the researcher to understand the error introduced and evaluate the appropriateness of the outputs. The reason for the missing data must be considered: while these tools are valuable for data that are missing at random, they are not suitable to systematic gaps such as signals below limit of detection.[48][26]

2.2 Methods

Let $X_i(t)$ represent the measurement of the i^{th} atmospheric feature at time t in data set X. At certain times ($t \in T_g$) the data from the i^{th} feature are unavailable and must be imputed. The remaining data points collected at times $t \in T_a$ with available data for the i^{th} attribute provide the training set for estimating the missing measurements. These points are also used as a test set to assess imputation performance by leave-one-out error analysis, discussed in §2.2.4.

2.2.1 Univariate Methods

2.2.1.1 Measurement Mean

The most straightforward approach is to replace all missing data with the mean of that measurement:

$$X_i(t \in T_g) = \frac{1}{|T_a|} \sum_{T_a} X_i(t). \tag{2.1}$$

2.2.1.2 Interpolation

Linear interpolation to fill the value of $X_i(t')$ in a gap from t_α through t_β simply calculates the values of X_i along a line from $[t_\alpha, X_i(t_\alpha)]$ to $[t_\beta, X_i(t_\beta)]$:

$$X_i(t') = X_i(t_\alpha) + (t' - t_\alpha)\frac{X_i(t_\beta) - X_i(t_\alpha)}{t_\beta - t_\alpha}. \tag{2.2}$$

2.2.1.3 Diel Average

Substitution with the diel average is a common approach for filling missing data in ground measuremens. If a data point is missing at time t', its value is estimated from T'_a, the set of points in T_a that occur at the same time of day:

$$X_i(t') = \frac{1}{|T'_a|}\sum_{T'_a} X_i(t). \tag{2.3}$$

Several other univariate spectral methods have been applied to air quality measurements to take into account periodicity on multiple timescales.[41] However, these approaches are subject to the limiting assumption that the most relevant points for reconstruction occurred at similar times of day/week/year. In reality, air quality measurements exhibit significant variability that cannot be captured in reconstruction from the frequency domain alone.

2.2.2 Multivariate Methods

2.2.2.1 Preliminary processing

The multivariate imputation methods estimate a missing value $X_i(t')$ with guidance from simultaneous measurements of other features $\{X_j(t'), X_k(t'), ...\}$. For a given data set X and target feature i requiring imputation of missing points, many of the other variables measured are likely to be redundant with each other or irrelevant to calculation of X_i. Indiscriminate variable inclusion increases computational cost and reduces imputation quality, so all multivariate analyses described below are preceded by stepwise feature

selection. This dimensionality reduction heuristic identifies a subset F of the available predictor variables whose inclusion minimizes mean square error in multilinear regression estimation of the target feature.[16]

The data for each feature is transformed by subtracting its mean ($\bar{X}_i$) and dividing by the standard deviation (σ_{X_i}) to produce z-scores with zero mean and unit standard deviation:

$$Z_i = \frac{X_i - \bar{X}_i}{\sigma_{X_i}}. \tag{2.4}$$

This normalization is necessary to remove anomalies arising from differences in arbitrary scales. Without normalization, the analysis response will be dominated by variables reported with large numerical values, e.g. temperature in Kelvin or pressure in torr $\gg$ glyoxal concentration in parts per billion.

Some data points will be more informative than others for reconstruction of a particular $X_i(t')$. The z-scores for the predictor features are used to identify the points in the training data set that are most similar to the point with missing data. These nearest neighbors are found by sorting the training points by their Euclidean distance to the missing point. The distance between data points collected at t' and $t \in T_g$ is given by

$$D[Z(t), Z(t')] = \sqrt{\sum_{f \in F} (Z_f(t') - Z_f(t))^2}. \tag{2.5}$$

2.2.2.2 Adaptive Average

Estimates for missing data can be obtained quite simply by a 'lazy learning' method that calculates the target feature using a distance-weighted average of nearby (similar) points. Given K neighbors at distances $\{D_1, D_2, ..., D_k, ..., D_K\}$ with values of the target feature $\{V_1, V_2, ..., V_k, ..., V_K\}$, the imputed value is calculated according to

$$X_i(t') \sim \frac{\sum_{k=1}^{K} D_k^{\rho} \cdot V_k}{\sum_{k=1}^{K} D_k^{\rho}}, \tag{2.6}$$

where the exponent ρ controls the weighting of the average with respect to distance. If $\rho = 0$, equation 2.6 reduces to the ordinary arithmatic mean of the neighbors' values. With $\rho = -1$ the average is weighted by inverse distance. Increasingly negative values for ρ increase the relative weights of the closer points. Tuning of the hyperparameters K and ρ is discussed in §2.3.3 and §2.3.4 respectively.

2.2.2.3 Regression Learning

Values for missing data can be imputed by training a regression learner on relevant subsets of the training data. Estimates are provided by a binary tree fit to nearest neighbors using k-fold cross-validation.

2.2.2.4 Artificial Neural Networks

Artificial neural networks (NN) are trained on relevant subsets to specialize in predicting a target feature based on the available measurements. The number of nodes in the input layer is equivalent to the number of selected features ($|F|$) and each receives z-scores from a predictor. The architecture of the hidden layer(s) may be tuned for particular conditions; for the data sets studied here, a single hidden layer was sufficient to perform effective dimensionality reduction and feature learning with low computational cost. A single output node corresponds to the target feature for imputation. The networks are trained by Levenberg-Marquardt backpropagation, which iteratively adjusts the weights and biases for each node to minimize error in the predictions of the target feature.[40][23] Evaluation of varying NN architectures is discussed in §2.3.5.

Since the backpropogation proceeds by stochastic descent methods, the network may train differently each time. A multiple imputation scheme can account for this by performing statistical analysis on many training/query replicates for the same points. Between each iteration, the neural network is reinitialized with random weights and biases, retrained, and queried for a single prediction. Methods for determining the appropriate cutoff for

NN replicates are discussed in §2.3.6.

2.2.3 Data Sets

2.2.3.1 Ground Measurements from GoAmazon

Data for assessment and testing on ground data were collected throughout the dry season during the Green Ocean Amazon 2014/15 field campaign.[36] Dozens of separate instruments were collocated at the T3 field site, 60 km to the west of Manaus, Brazil (3Â°12’47.82"S, 60Â°35’55.32"W). The data used contains ∼4,700 timepoints at 15-minute resolution with measurements of 20 chemical and 5 meteorological features.

2.2.3.2 Flight Measurements from SONGNEX

Flights for the Shale Oil and Natural Gas Nexus (SONGNEX, [42]) campaign were conducted during April-May of 2015 in the NOAA WP-3D research aircraft. Imputation methods are tested on the 2015.04.13 flight north of Denver, CO, USA. The data used contains ∼18,000 timepoints at 1-Hz resolution with measurements of 35 chemical and 5 meteorological features.

2.2.4 Performance testing

The performance of each imputation method was assessed by leave-one-out error analysis. All data were masked within a time window around each point to be imputed, since sampling in the immediate vicinity would effectively approximate interpolation. For testing on the GoAmazon ground data, a 24-hour window centered on each data point was left out so that estimates were based on data from other days. For the SONGNEX flight data a 5-minute window masked, corresponding to approximately 36 km at the median ground speed of 120 km/hr.

Details of feature selection and hyperparemeter tuning are discussed for formaldehyde (HCHO) measurements in both field campaign data sets. Results for additional chemical species are presented in §2.3.7.

2.3 Results

2.3.1 Feature Selection

The stepwise feature selection identifed subsets of 15 measurements that are efficient predictors for HCHO, including boh chemical and meteorological features. In the GoAmazon data set, the strongest predictors were ozone, methyl vinyl ketone + methacrolein, temperature, relative humidity, pressure, isoprene, and shortwave irradiation. In the SONGNEX data set, the strongest predictors were ozone, carbon monoxide, acrylonitrile, acetonitrile, nitrogen dioxide, and dew point temperature.

2.3.2 Neighbor selection

The nearest neighbor search described in equation 2.5 identifies points for the adaptive average with conditions most similar to the points being imputed.

2.3.2.1 GoAmazon examples (ground)

The top half (panels *a* and *b*) shows estimation of a data point from around noon local time on 19-September 2014, which exhibited relatively low HCHO. Panel *a* shows the HCHO measurement timeseries for the entire dry season. Panel *b* shows the same data presented with time of day as the x-axis, with the corresponding points marked. The masked point is shown in red, while the points that are used in calculation of the diel average are highlighted in yellow. Due to lower-than-typical HCHO concentrations during the day of interest, most of the values included in the diel average are higher (some more than double). Blue markers indicate nearest neighbors, showing that the most similar

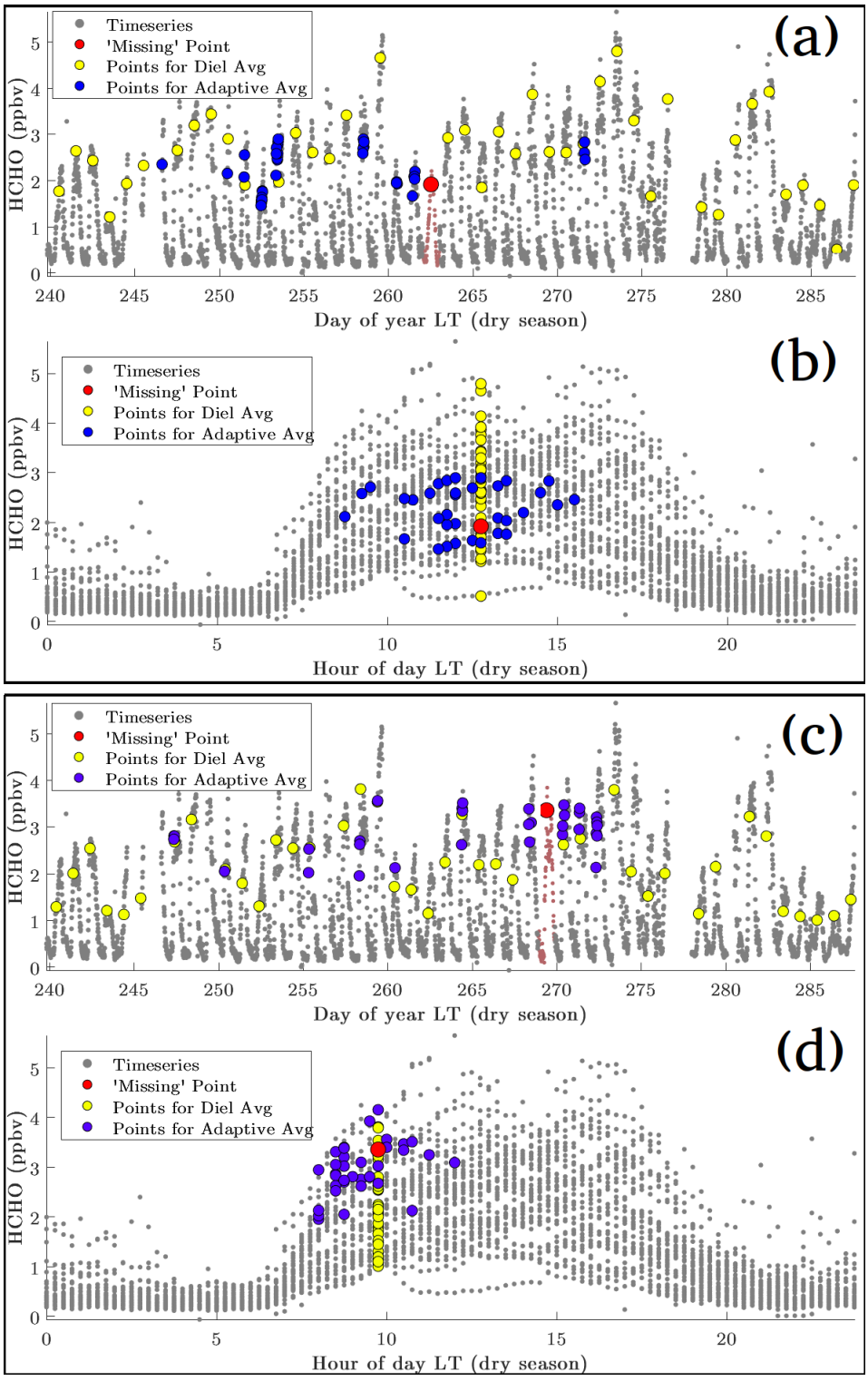

Figure 2.1: Data points used for estimation of 'missing' HCHO on 19-Sept 2014 (*a* & *b*) and 26-Sept 2014 (*c* & *d*) during the GoAmazon campaign. The top panel for each day shows a masked point to be imputed (red marker), the data points that are used for the diel average (yellow markers, at the same time each day), and the times with similar conditions identified for the adaptive average (blue markers, from nearest-neighbor search). The top panel for each day shows the timeseries over the course of the dry season, while the bottom panel for each shows the same data plotted is a function of time of day.

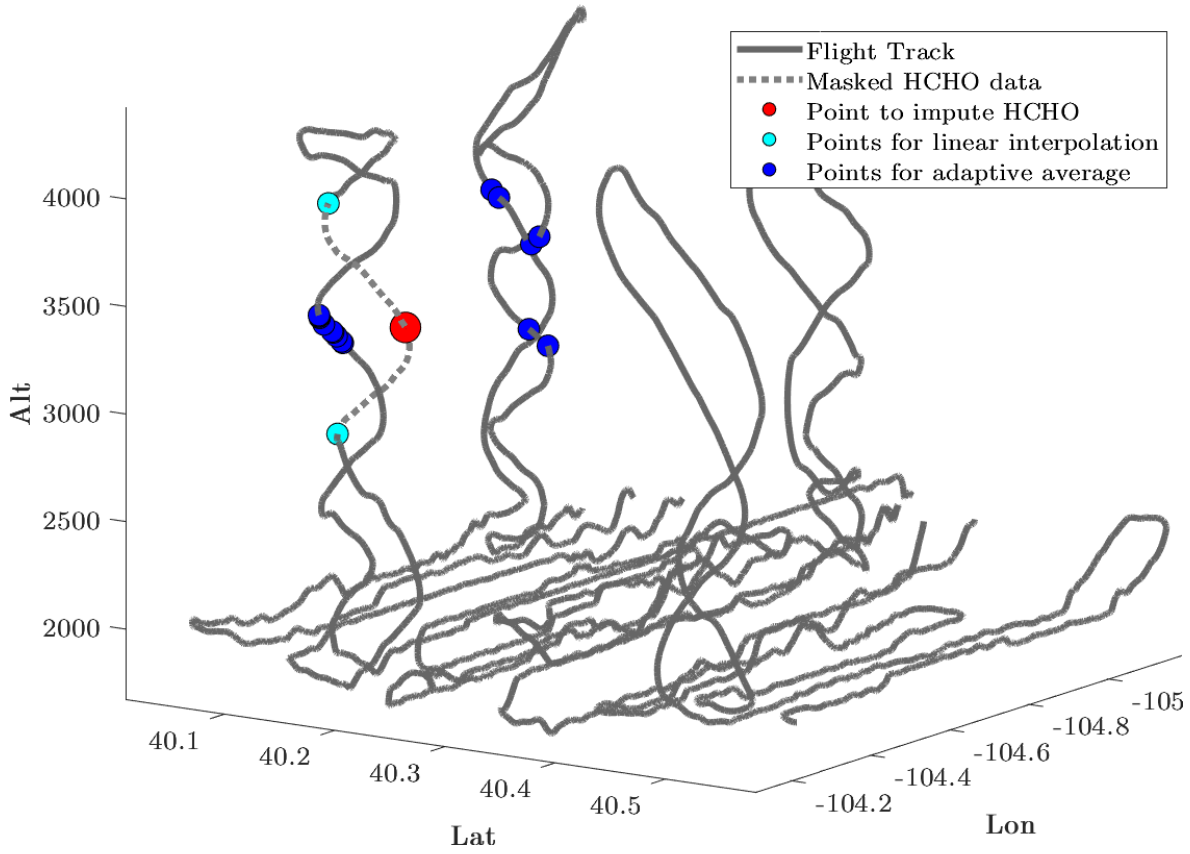

Figure 2.2: Data points used for estimation of a 'missing' HCHO measurement in the free troposphere during the SONGNEX research flight. The red marker indicates the masked point to be imputed. The cyan markers show the points that would be used for imputation by linear interpolation, and the blue markers show points with similar conditions identified for the adaptive average.

conditions were found in a 3 week window around the masked point (see *a*) and occurred throughout the day from 9:00 h - 15:50 h LT (see *b*).

The bottom half (panels *c* and *d*) shows estimation of a data point from the morning of 26-September 2014, which exhibited higher-than-average HCHO. The most similar conditions were identified during the mornings of other days with elevated HCHO.

2.3.2.2 SONGNEX examples (flight)

Figure 2.2 compares the datapoints that linear interpolation and adaptive averaging take into account for estimation of a masked point (red marker) during an ascent through the

free troposphere. The linear interpolation calculates estimates based on the two cyan points 500 km above and below the point to be imputed. The blue markers indicate nearest neighbors identified for adaptive averaging (i.e. the timepoints corresponding to the 12 shortest distances calculated from equation 2.5). Note that these most similar conditions were identified on the subsequent descent near the same altitude, and on the spiral that was closest, both spatially and temporally.

Figure 2.3 shows which data points were selected for estimation of a masked point (red marker) during a biomass burning plume sampled in the planetary boundary layer. The masked HCHO data and the points found as nearest neighbors were plotted on top of ozone measurements, since ozone is the strongest predictor for HCHO in the SONGNEX data set (see §2.3.1). The nearest neighbors were identified downwind of the source, near the edges of the plume with similar extent of photochemical processing.

2.3.3 Number of Training Points

For all of the multivariate methods, the number of training points to be used for the average or learning methods must be selected.

Figure 2.4 shows how adjusting the number of nearest neighbors impacts the performance of the adaptive averaging method. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. Mean average error (top) and method bias (middle) is quantified for the range of HCHO concentrations in each data set. Histograms showing the HCHO variation (measurements) are included at the bottom. Performance is shown as a function of concentration, since aggregate statistics (such as overall r^2) do not capture range-dependent error and can be misleading.[59][60]

Using anywhere between 5 and 100 nearest neighbors produces reasonable results. Increasing the number of points included in the average beyond 100 degrades performance by including irrelevant points that bias the estimate toward the mean. The error from this effect is most pronounced at the extremes of the meauserement range. If all data points

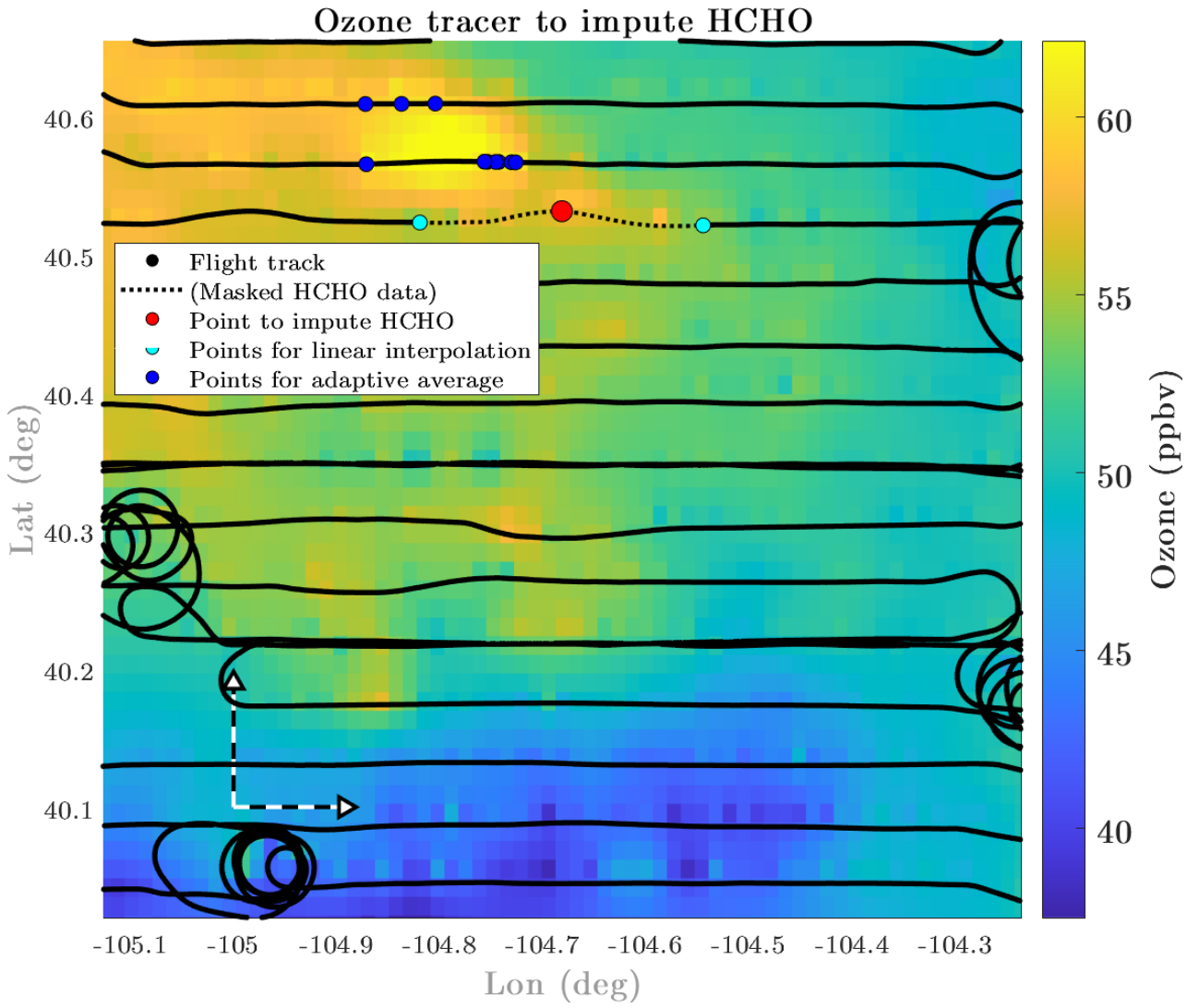

Figure 2.3: Data points used for estimation of a 'missing' HCHO measurement from the planetary boundary layer during the SONGNEX research flight. The red marker indicates the masked point to be imputed. The cyan markers show the points that would be used for imputation by linear interpolation, and the blue markers show points with similar conditions identified for the adaptive average. The background shows the (smoothed) spatial distribution of ozone in the planetary boundary layer, which was the strongest predictor for HCHO.

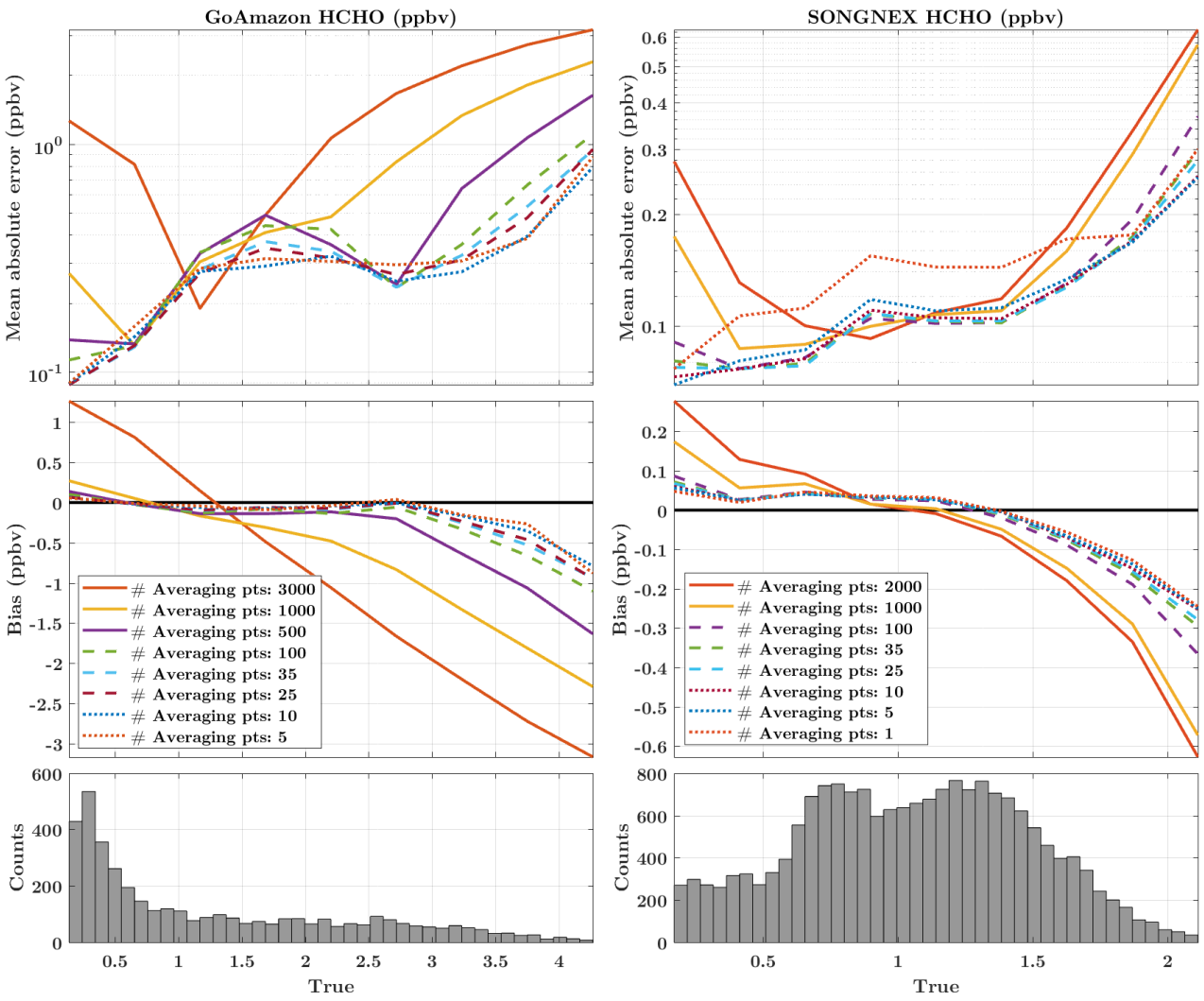

Figure 2.4: Plots showing error analysis used for tuning the number of points to be included in the adaptive average. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set.

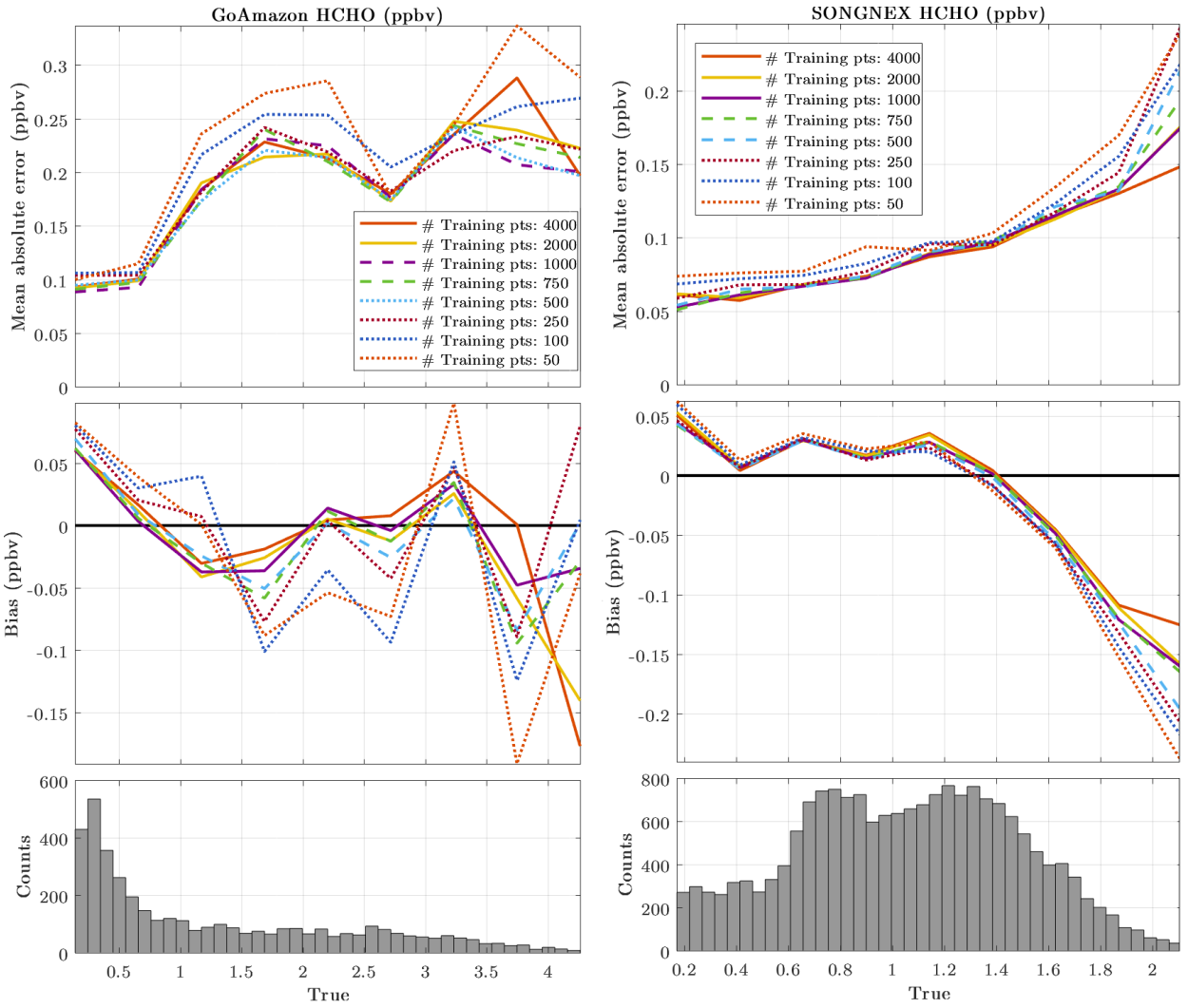

Figure 2.5: Plots showing error analysis used for tuning the number of points to train the neural network. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set.

are included (in an unweighted average), the adaptive average becomes the measurement mean.

Figures 2.5 and 2.6 show impacts of changing the size of the training set for the NN and regression learner, respectively. The ANN performance increses with the size of the training set, and including fewer than 250 points markedly reduces performance. The regression learner performs relatively well regardless of the size of the set, except that estimates for

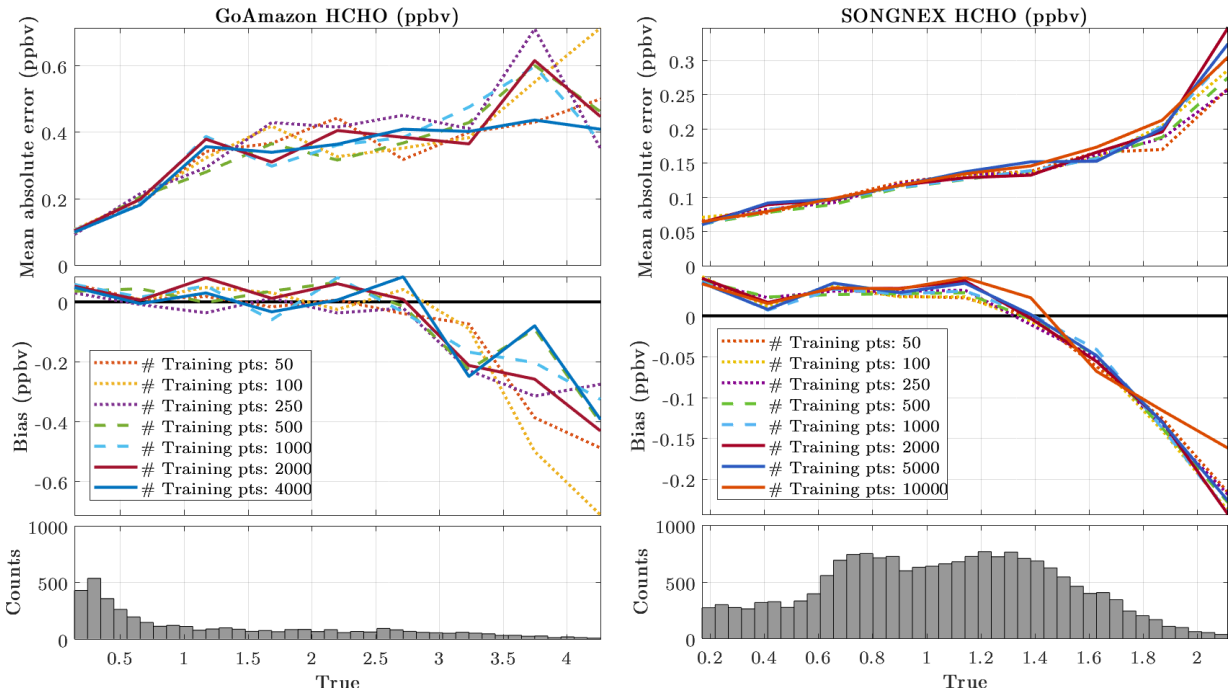

Figure 2.6: Plots showing error analysis used for tuning the number of points to train the regression learner. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set.

high concentrations were biased low if fewer than 250 data points were included in the GoAmazon data set.

2.3.4 Adaptive Average Weighting

The hyperparameter ρ in equation 2.6 controls the weighting of points in the adaptive average depending on their distance. The effect on neighbor significance is shown for an example point in figure 2.7, which traces the stepwise calculation of the cumulative mean with varying values of ρ. The more negative the value of ρ, the less responsive the mean to the addition of subsequent points. For points whose neighbors' values are distributed randomly around the mean, the same average is quickly attained regardless of weighting scheme. The impact of ρ on overall imputation performance is quantified in figure 2.8. Estimates from the GoAmazon data set were relatively insensitive to ρ except

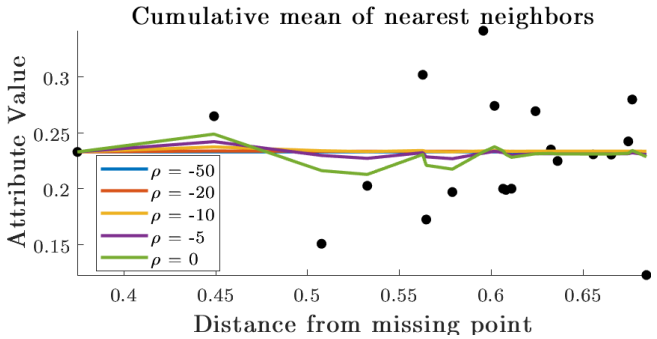

Figure 2.7: Distance values of nearest neighbors as a function of their distance from the masked point (calculated from the z-scores as in equation 2.5. The cumulative mean is shown with the stepwise addition of each point.

for imputation of higher concentrations, where $\rho \leq -10$ produces slightly better results. For the SONGNEX data set $-10 < \rho < 0$ produced the lowest mean absolute error, and bias was not significantly impacted by any value of ρ.

2.3.5 Neural Network Architecture

For the data sets tested, the performance of the NN was very robust to configuration of the hidden layers. Only a few nodes were needed for effective pattern recognition training. Single and multilayer configurations were tested, with results shown in figure 2.9. Good performance was attained with a single hidden layer containing 5 nodes, as shown in figure 2.10.

2.3.6 Neural Network Replicates

Because NNs do not train deterministically, different results may be obtained with each training. Consequently, a multiple imputation scheme was adopted. NNs were repeatedly reinitialized and retrained, then the mean is calculated from the collection of replicate outputs. A minimum of 20 replicates are run to ensure sufficient sample size. The retraining continues until 6 replicates in a row do not change the mean by more than 0.5%, at which point the cycle is halted. Figure 2.11 shows the results of 100 NN replicates (more

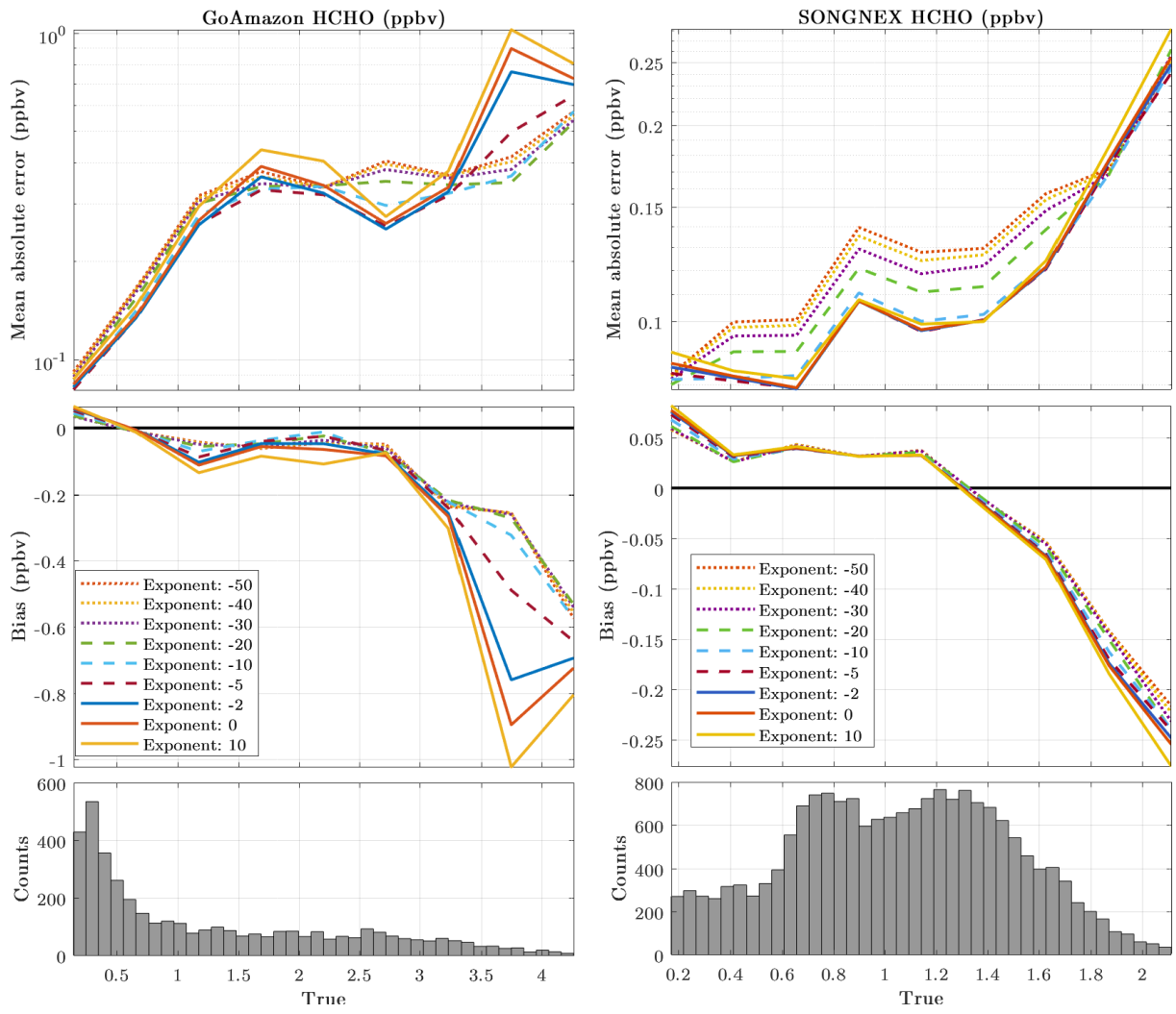

Figure 2.8: Plots showing error analysis used for tuning the number hyperparameter ρ, the distance exponenent in the weighting function. The left plots show results from the GoAmazon data, and the right plots show results from SONGNEX data. The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set.

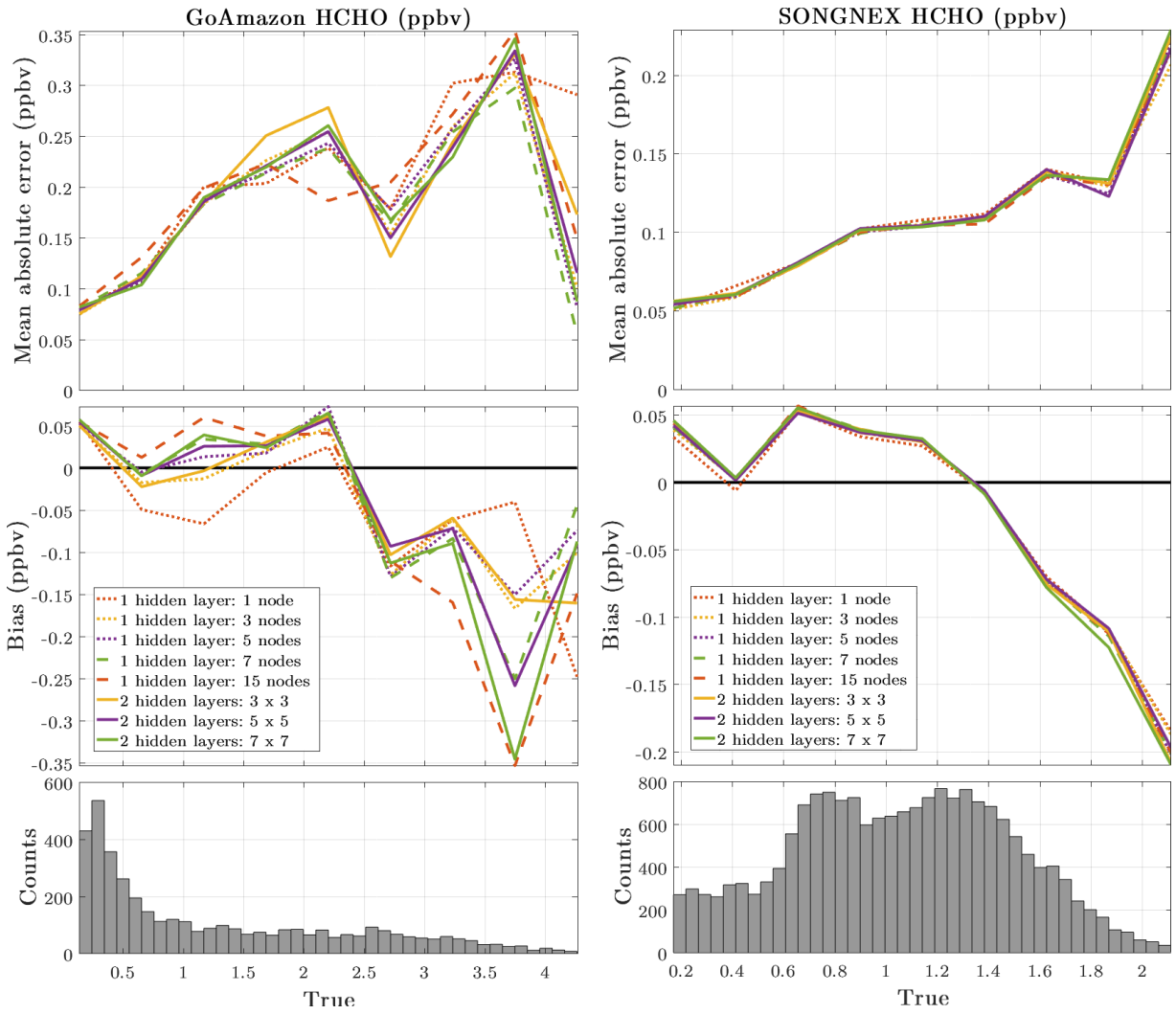

Figure 2.9: Plots showing error analysis used for testing various NN configurations on the GoAmazon data set (left) and SONGNEX data set (right). The top plots show mean absolute error, the middle plots show bias, and the bottom plot shows the distribution of HCHO concentrations in each data set.

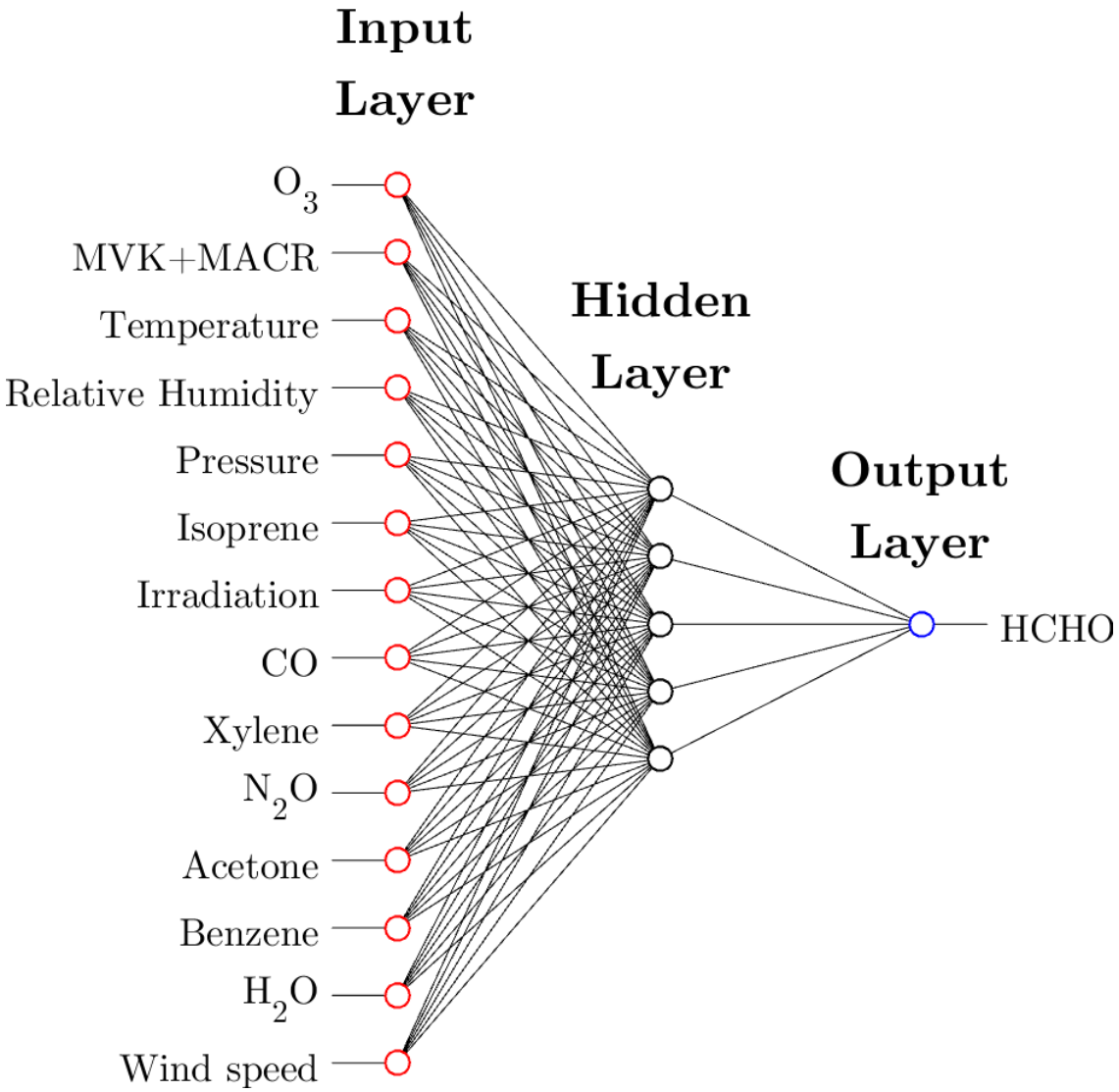

Figure 2.10: Diagram of the NN configuration generated for imputing HCHO in the GoAmazon data set.

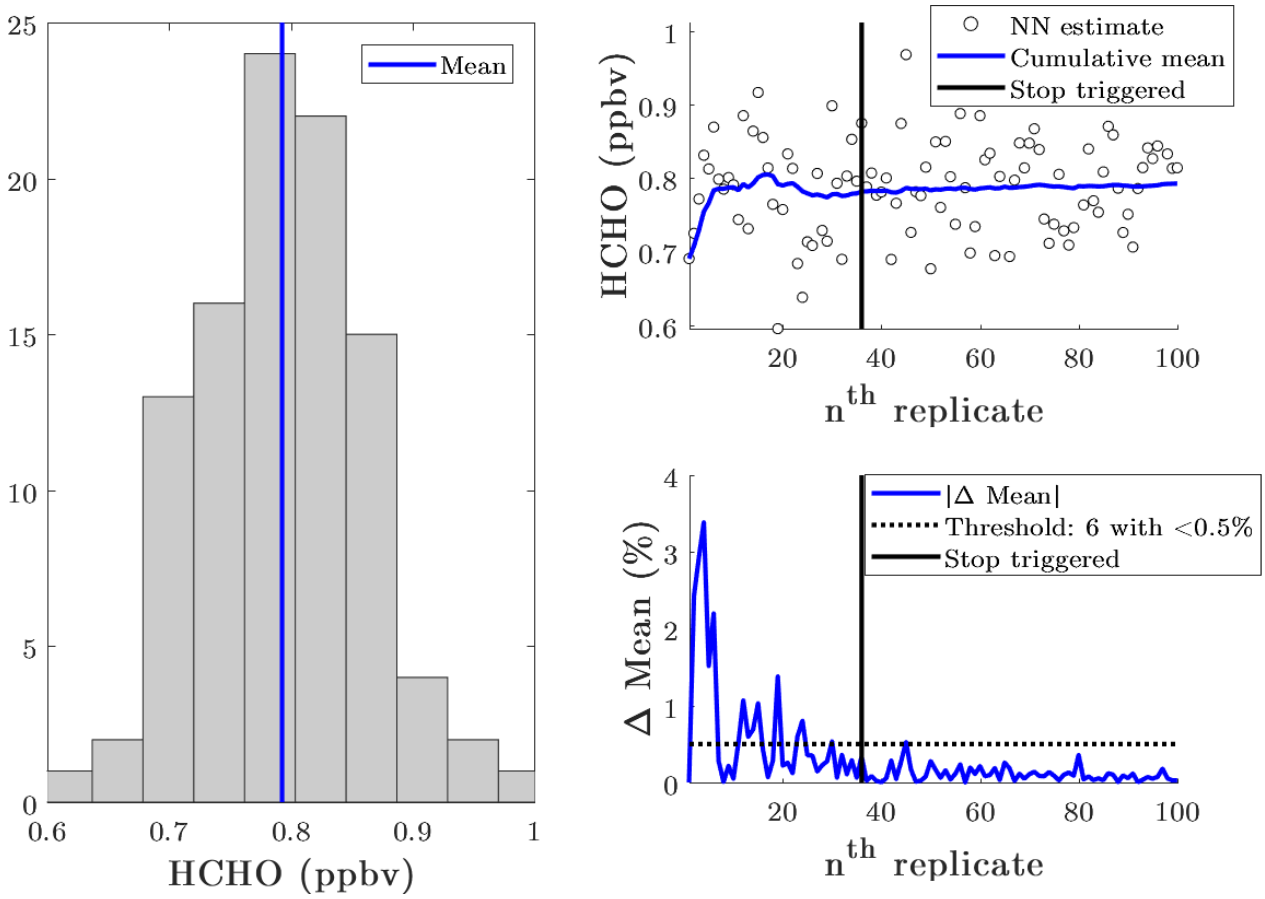

Figure 2.11: Statistical analysis of NN outputs for 100 training replicates for the same point. The histogram on the left shows the distribution of estimations. The upper right figure shows the value of each estimate and stepwie calculation of the mean (blue trace). The bottom right figure shows the change in mean with each successive addition (blue trace), along with the threshold for stable mean (dotted black line) and the replicate at which the criteria are met (solid black line).

than necessary). The left panel histogram displayes the distribution of NN estimates for imputation of a given point. Each estimate is shown in the upper right hand panel (black circles) along with the mean including all estimates up to a given replicate (blue trace). The blue trace in the lower right panel shows the change in the mean with the addition of each successive estimate. The horizontal dotted line shows the threshhold selected for halting (0.75%), and the vertical black line indicates the replicate at which the criteria were met.

For the majority (65%) of data points, a stable mean triggered NN halting as soon as the minimum number of replicates was completed. Figure 2.12 shows the frequency of NN

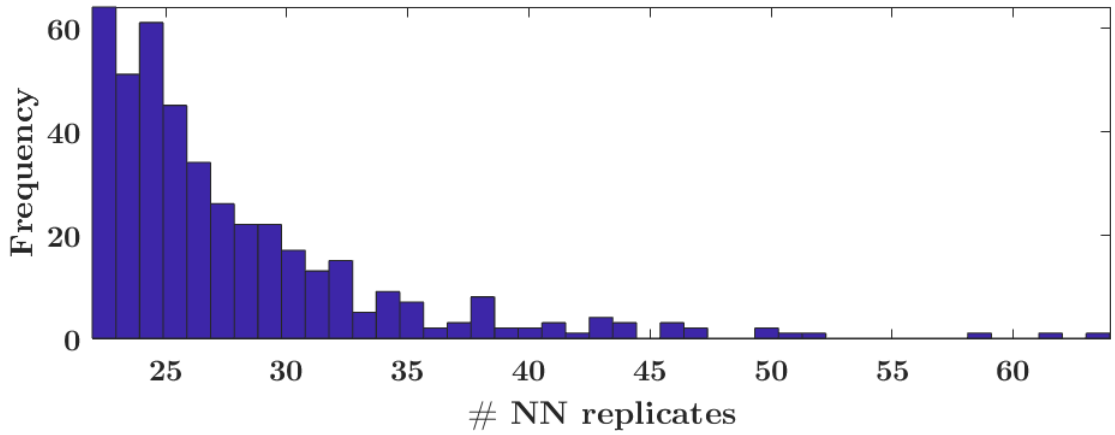

Figure 2.12: Number of NN replicates required to meet halting conditions (<0.5% change in mean with the inclusion of 6 successive replicates) in the GoAmazon data set. The 65% that halted at the minimum number of estimates (20) are not shown; these are statistics for the 35% that required more replicates.

training cycles for the GoAmazon data set that required more than the minimum number of replicates to achieve the halting criteria (i.e. the other 35% of data points). More than 90% of points required ≤30 replicates.

2.3.7 Intercomparison

Figure 2.13 shows pointwise intercomparison of the imputation methods. A data point whose reconstructed value matches the true masked value falls on the 1:1 diagonal. The further from the diagonal, the more that a point was under/over predicted. The top plot shows comparison for HCHO imputed during the GoAmazon campaign (with comparison to diel average) and the bottom plot hows the comparison for HCHO during the SONGNEX flight (compared to linear interpolation over the ∼36 km synthetic gaps).

The intercomparison between all methods for imputing HCHO is displayed in figure 2.14 (GoAmazon on the left, SONGNEX on the right). The performance is quantified as a function of the feature range to display variation in performance at varying concentrations. The top panels show the mean absolute error, and the middle panels show the bias. The bottom panels show the distribution of HCHO in each data set, for reference.

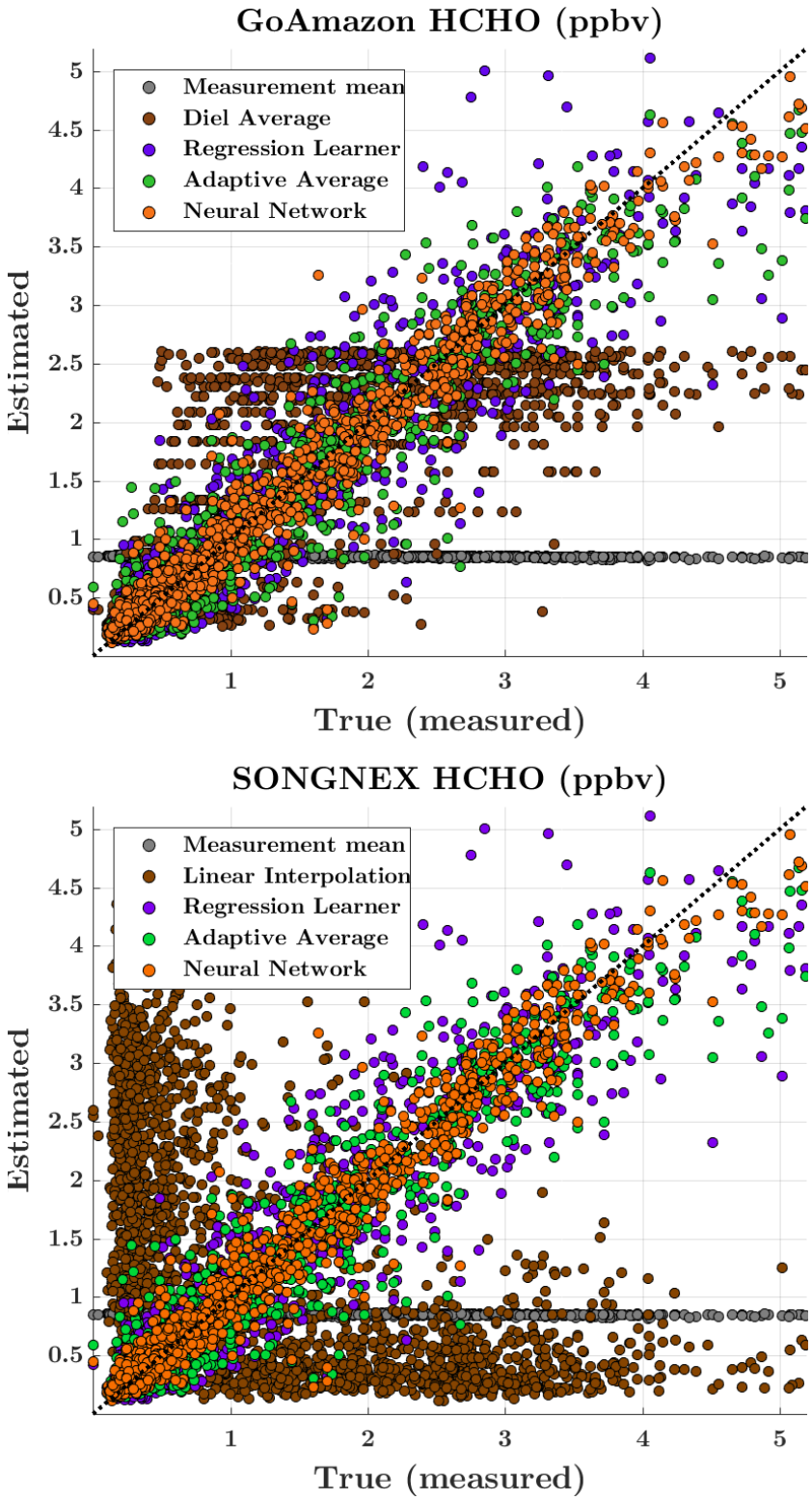

Figure 2.13: Pointwise intercomparison of imputation methods with GoAmazon ground measurements on the top and SONGNEX flight measurements on the bottom. A data point whose reconstructed value matches the true masked value falls on the 1:1 diagonal. The further from the diagonal, the more that a point was under/over predicted (to the right and to the left, respectively).

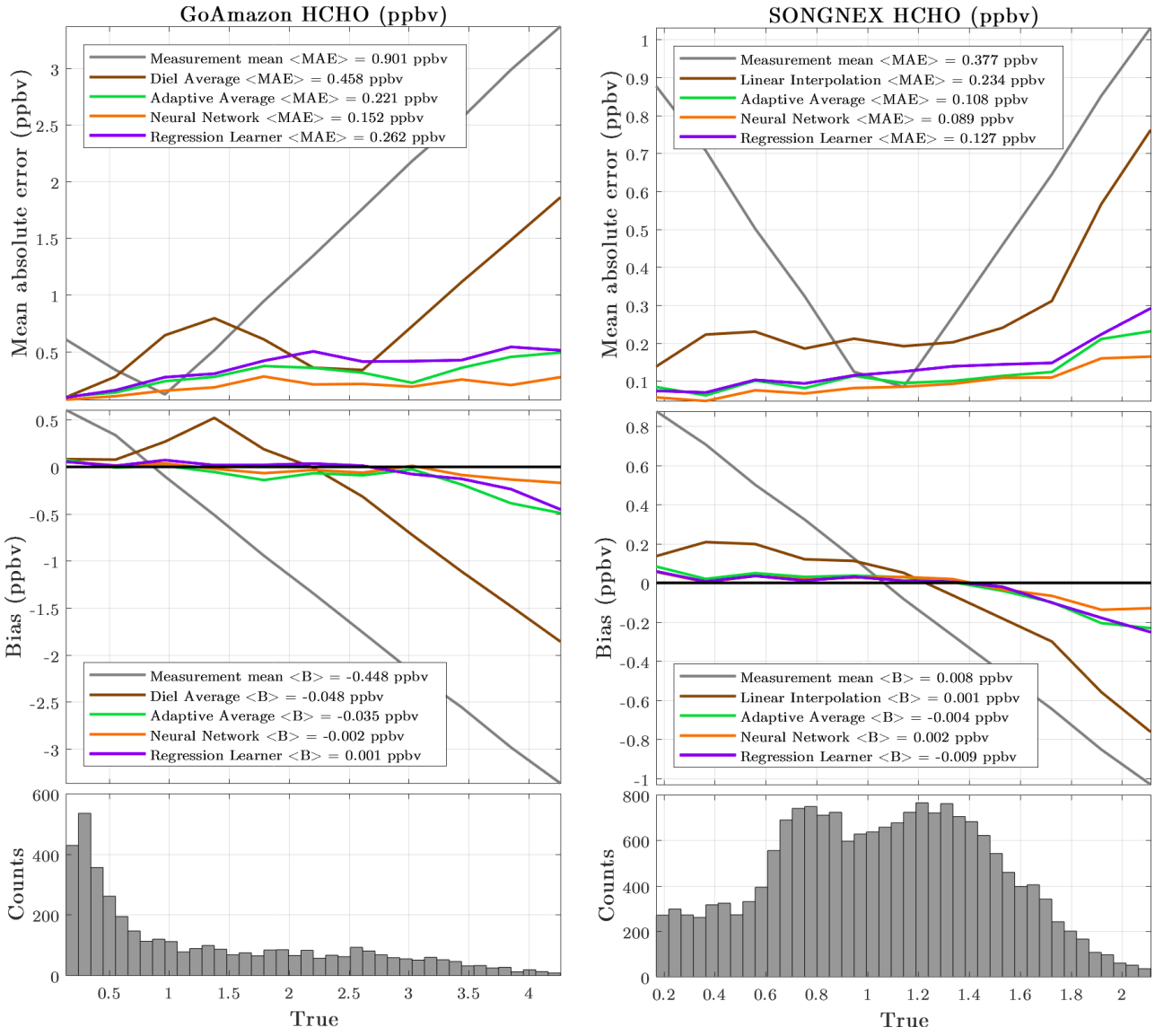

Figure 2.14: Intercomparison of imputation methods for HCHO field data (GoAmazon left, SONGNEX right). The top panels show mean absolute error (MAE) at varying concentrations, with with the pointwise average MAE for each method in the legend text. The middle panels show the bias as a function of concentration, with the average bias in the legend text. The bottom panels show the distribution of HCHO in each data set, for reference.

The measurement mean (gray trace) performs poorly as a substitution for all points. Naturally, the bias and error are most exaggerated at the extremes of the concentration range.

The diel average (brown trace, GoAmazon at left) exhibits more error than the multivariate methods and significant bias (up to -40%). The bias at high concentrations is problematic when modeling days with pollution events, since the diel average introduces a strong negative bias (see figure 2.15 highlighting this underestimation for isoprene, a biogenic precursor for photochemistry that leads to the production of ozone and secondary aerosol).

The linear interpolation (Figure 2.14 brown trace, SONGNEX at right) performs better than the measurement mean. The 5-minute mask means that any feature $<$30 km in scale will be entirely missed by the linear interpolation, so overlooked enhancements result in negative bias at higher concentrations.

The multivariate methods produce significantly better results in both data sets, with mean absolute error less than half of the univariate methods. Additionally, bias is markedly reduced across the range of concentrations ($\leq$15%). The NN performed the best in both data sets, followed the adaptive average.

Figure 2.16 shows the performance of each imputation method on the ozone measurements during both field campaigns. Figure 2.17 shows performance for imputating the number concentration of fine particulate matter (optical diameter 0.004-1.0 μm) during the SONGNEX flight.

2.4 Conclusions

A variety of machine learning tools (lazy learners, neural networks, regression learners) are flexible and effective for filling gaps in atmospheric data sets. Which measurement types should be used as inputs can be ascertained by stepwise feature selection algorithms.

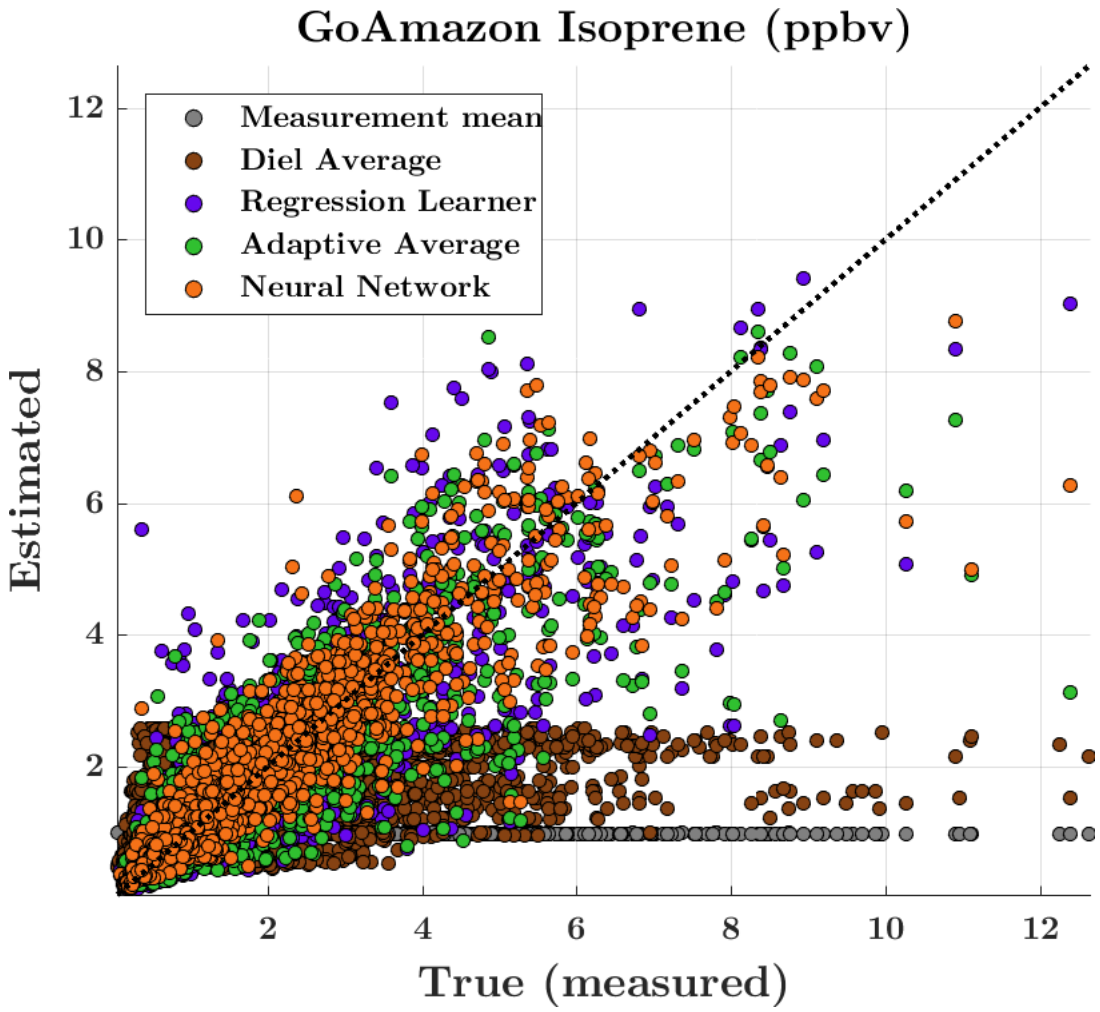

Figure 2.15: Intercomparison of adaptive average and diel average imputation estimates for isoprene during GoAmazon. A data point whose reconstructed value matches the true masked value falls on the 1:1 diagonal. The further from the diagonal, the more that a point was under/over predicted (to the right and to the left, respectively).

Performance was quantified by leave-one-out error estimation on a variety of measurement types (HCHO, ozone, isoprene, particulate matter) for ground and flight data. For all intercomparisons, the specialized NN produced the best estimates for 'missing' data points, across a wide variety of features and conditions. The adaptive average performs nearly as well, and requires no replicates. The barrier to adoption of this method is low, since the nearest neighbor search on z-scores is simple in concept and implementation.

While a wide range of values for each paramater was tested here for the sake of thorough

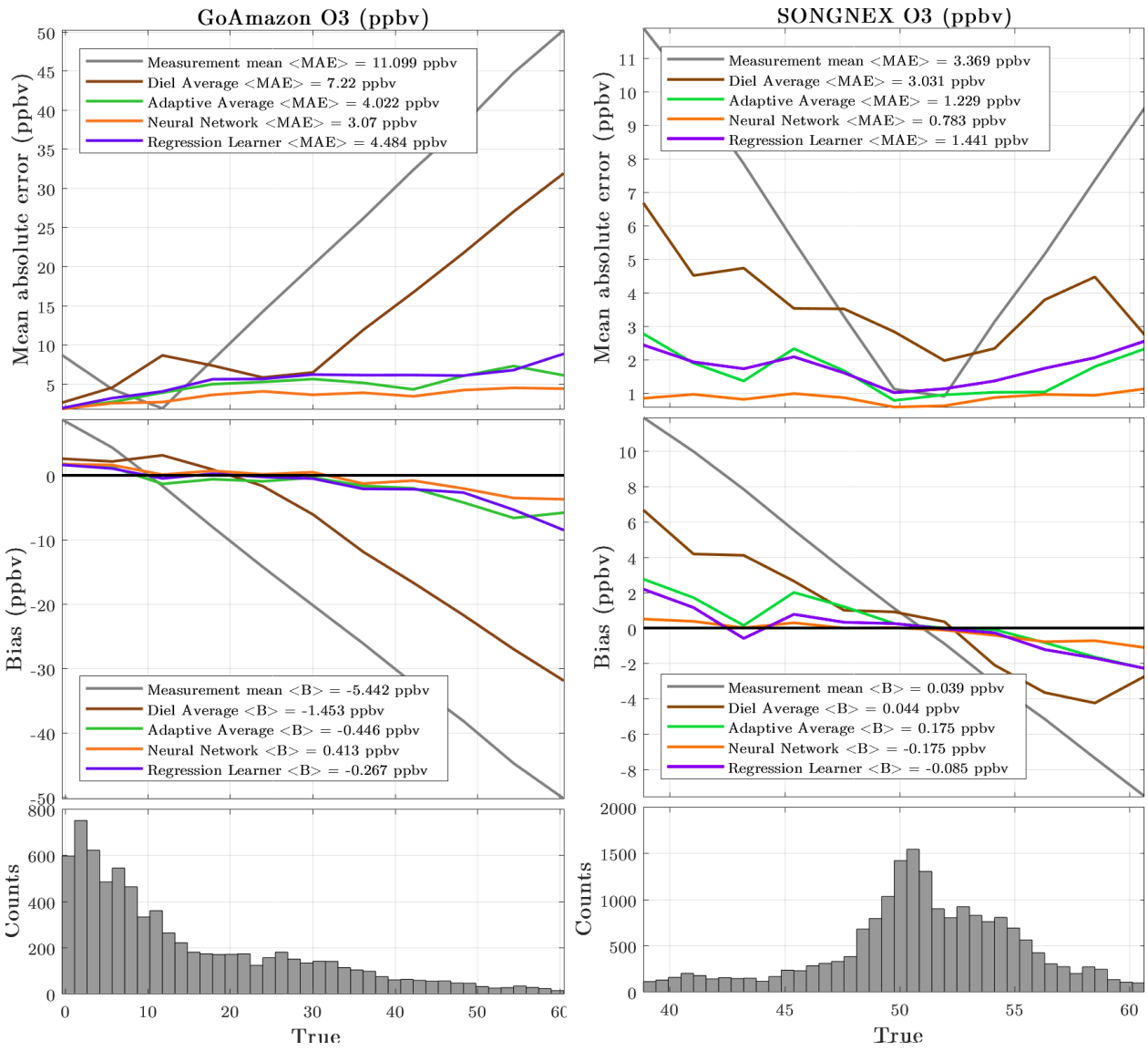

Figure 2.16: Intercomparison of imputation methods for ozone field data (GoAmazon left, SONGNEX right). The top panels show mean absolute error (MAE) at varying concentrations, with with the pointwise average MAE for each method in the legend text. The middle panels show the bias as a function of concentration, with the average bias in the legend text. The bottom panels show the distribution of ozone in each data set, for reference.

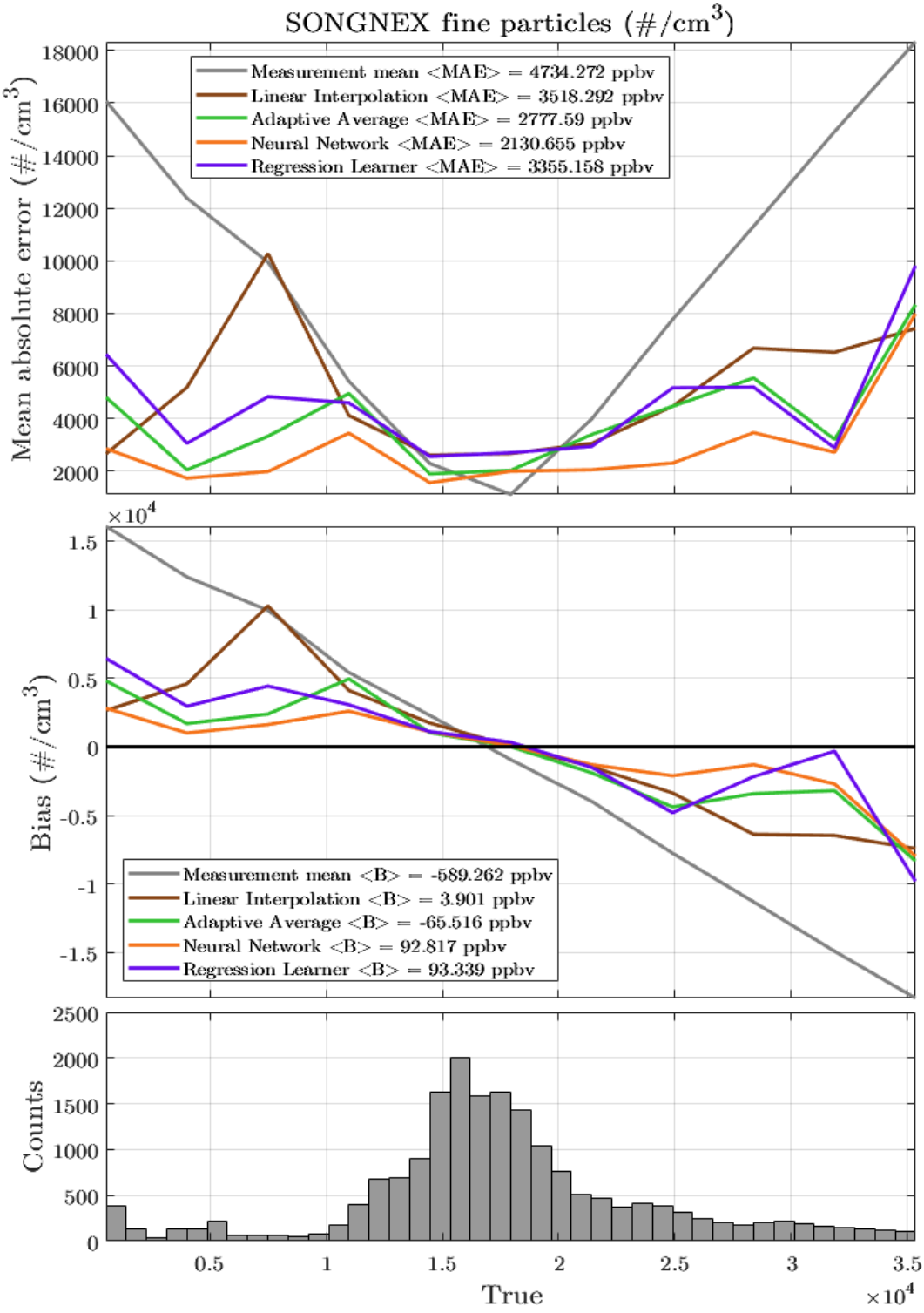

Figure 2.17: Intercomparison of imputation methods for fine particulate matter (optical diameter 0.004-1.0 μm). The top panels show mean absolute error (MAE) at varying concentrations, with with the pointwise average MAE for each method in the legend text. The middle panels show the bias as a function of concentration, with the average bias in the legend text. The bottom panels show the distribution of ozone in each data set, for reference.

quantification, such analysis is not necessary for initial application to new data sets. For most of the parameters, a wide range of values resulted in accurate imputation. For NN imputation, a configuration with a single hidden layer containing 5 nodes should provide a good starting point. For the adaptive average, there is some redundancy between the number of neighbors included in the average and the ρ exponent (e.g. for a given weighting function, a similar result can be acheived with more neighbors and a more negative exponent). Consequently, initial performance can be tested with $\rho = 0$ over varying number of neighbors. For the data sets tested here, averages using 10-35 neighbors provided good results, however this will depend on each data set and its time resolution.

Upon application to a new data set, the imputation performance should be assessed by some method of cross-validation on the training points such as the leave-one-out error analysis described in §2.2.4. This is necessary to provide an estimate of the error and bias to expect from imputed values, in order to ascertain the overall validity of a given imputation method for that particular data set. Additionally, the error distribution for each target feature could be used to produce a bias-corrected jackknife estimate for imputed values.

2.5 Future work

While presented in the context of imputation, these methods can be applied to combining data from collocated instruments onto the same timebase. Timepoints containing data from the feature sampled less frequently serve as the training set to 'impute' the slower timeseries onto the faster timebase. This is likely to produce better results than linear interpolation for timebase matching, especially if multiple measurements already exist on the faster timebase.

With some light modifications, the multivariate methods described in this paper could be tested for forcasting atmospheric chemistry (instead of imputation). Nearest neighbors would be identified from similar conditions as above, however the learners would be

trained to predict the *derivative* the target feature instead of its absolute value. The methods would thus predict the chemical gradient based on the chemical gradients of the nearest neighbors. Preliminary stepwise feature selection would be carried out with the same predictors, with the feature derivative as the response variable. Outside of the context of imputation, the value of the target feature can be included as a method input.

Estimates for all of the gaps in a data set can be calculated simultaneously using an iterative expectation maximization technique popularized by the Netflix prize for new recommender systems.[31][22] These methods take advantage of a singular vector decomposition (SVD), a dimensionality reduction technique that produces the best low-rank approximation of any input matrix. Missing data to be imputed are filled in with any rough estimate (e.g. measurement mean) and the decomposition is calculated. Then the SVD is used update the initial estimates for the missing point; this process repeats until estimates converge. This technique is effective even for extremely sparse (e.g. 99% missing) matrices, and has great potential for atmospheric measurements.

2.6 Acknowledgements

The author acknowledges guidence and insight from Amir H. Assadi, who initially pointed out that the nearest neighbors identified for the adaptive average could also be used to train neural networks. Thanks to Abigail G. Thayer for suggesting the network replicate halting criteria based on a threshold for change in the cumulative mean.

3 IDENTIFICATION OF PLANETARY BOUNDARY LAYER HEIGHT FROM SELF-SIMILARITY IN VERTICALLY-RESOLVED MEASUREMENTS

ABSTRACT: The height of the planetary boundary layer (PBL) significantly impacts pollutant concentrations near the ground by affecting the extent of vertical mixing. The structure of the atmosphere changes over the course of the day, and the dynamic PBL height must be measured or estimated for many chemical and meteorological models and forecasts. This work details a new method determining the height of the PBL from in-situ measurements passing through both layers (e.g. radiosonde or flight data).

The PBL and free troposphere (FT) are largely decoupled, so variation between the PBL and FT is greater than variation within each layer for many physical/chemical quantities. Consequently, techniques from cluster analysis can be used to identify the height of the PBL based on self-similarity in vertically-resolved measurements. This is accomplished by identifying the global maximum in cluster evaluation indices calculated for hypothetical partitions at each height measured during a given profile.

Evaluation using silhouette values and Calinski-Harabasz indices both yield clear maxima in agreement. Techniques and shortcomings of forward clustering algorithms (centroid-based, density-based, and Gaussian mixtures) are discussed.

This approach is demonstrated on flight data sets from the Pan-European Gas Aerosol Climate Interaction Study (PEGASOS 2013, zeppelin) and Shale Oil and Natural Gas Nexus (SONGNEX 2015, WP-3D Orion) field campaigns. This analysis requires no parameterization and is shown to be robustly applicable to multiple classes of measurements (metorological, trace gases, volatile organic compounds, reactivity).

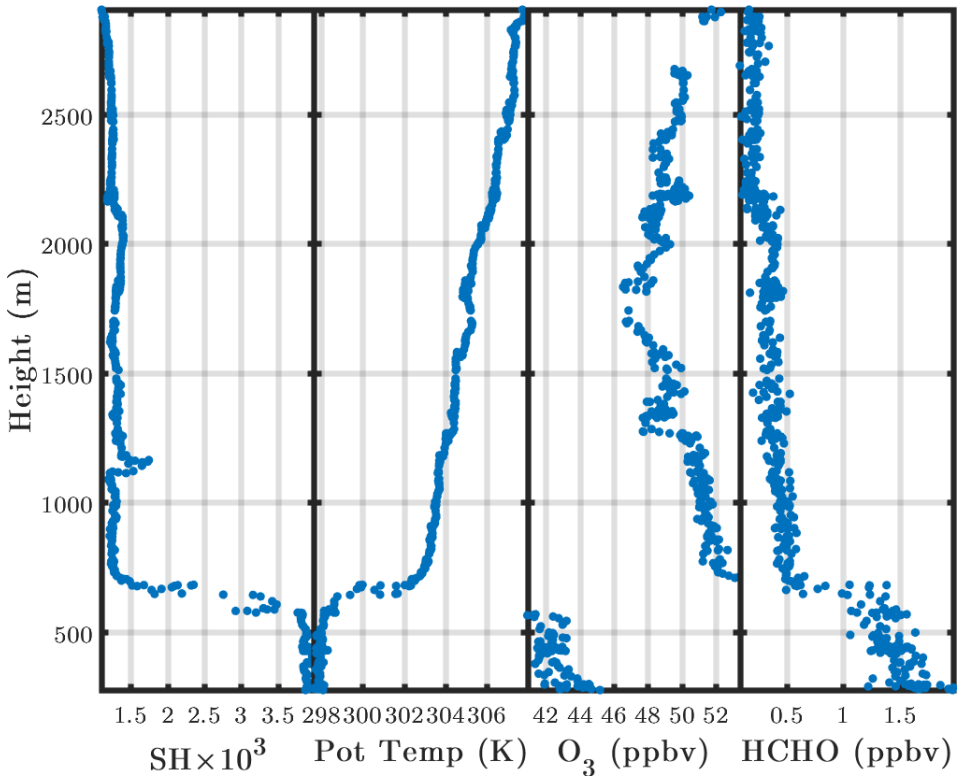

Figure 3.1: Measurements from the first SONGNEX vertical profile. Specific humidity (SH), potential temperature (θ), ozone (O_3), and formaldehyde (HCHO) exhibit discontinuities near $z' \sim 650$ m.

3.1 Introduction

The lowest layer of the atmosphere is the planetary boundary layer (PBL), often topped by a capping inversion layer, and above that the free troposphere (FT). Since vertical mixing occurs mostly within the PBL, its height (z') plays a key role in determining pollutant concentration near the surface. Knowledge of the diurnal PBL height is useful for many air quality models, and must be modeled or measured. Consequently, a common motif in aircraft campaigns is to periodically fly vertical profiles (e.g. from 500 m to 2000 m and back) in order to map out the structure of the atmosphere. Current methods for identifying the boundary layer height from in-situ data rely on identifying particular features in the profiles such as: gradients in specific humidity (SH), potential temperature (θ), or derived Richardson number.[56][53]

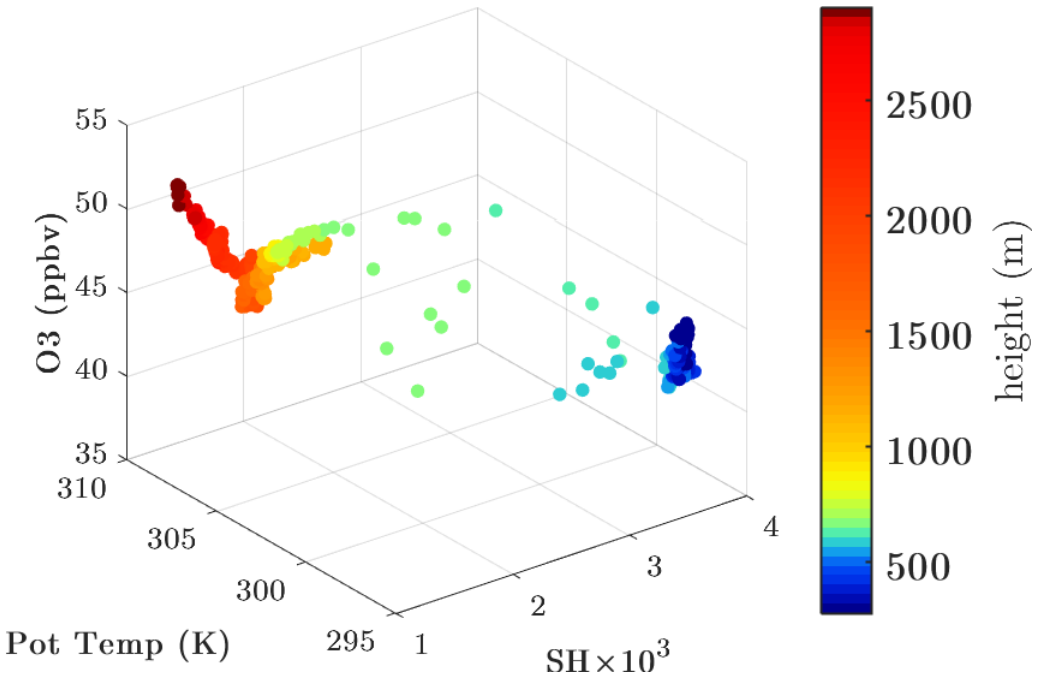

Figure 3.2: Measurements from the first SONGNEX vertical profile. Specific humidity (SH), potential temperature (θ), and ozone (O_3). The data are colored by height to highlight the separation between clusters near $z' \sim 650$ m.

An alternative paradigm is presented here, exploiting the self-similarity of data within a given atmospheric layer. Methods for cluster analysis and cluster evaluation can adeptly discern z' from the topology of vertically-resolved measurements. For illustration, figures 3.1 and 3.2 show typical data for a PBL→FT profile, from the first ascent of a SONGNEX research flight on 2015.04.13 (details regarding the measurements, platform, flight plans, and research goals are available elsewhere[42]).

Figure 3.1 shows the specific humidity (SH), potential temperature (θ), ozone (O_3), and formaldehyde (HCHO) from this first SONGNEX profile, where $z' \sim 650$ m. The left two panels show that the PBL is relatively moist with low θ, while the FT is drier with higher θ. The right two panels show corresponding discontinuities in mixing ratios that occur at the z'.

Figure 3.2 shows scatter plot of SH, θ and O_3 (the left three panels of 3.1) with height indicated by color. The cold moist air in the PBL clusters to the right side of the figure,

while the warm dry air in the FT clusters to the left side. The colorscale highlights the large margin between clusters in the data corresponding to a small height difference around z'

This two-cluster topology is a natural feature of data from multiple layers of the atmosphere, and techniques described here all seek to discern the height that best partitions the measurements into 'best' clusters. These methods requires no parameterization, and thus no subjective selection of criteria or threshold values. Artifacts within a layer arising from stratification, advection, or other odd vertical structures that confuse gradient searching algorithms generally result in shifts within a cluster, so the results are robust to such disruptions.

3.2 Theory

3.2.1 Data Preparation

Any subset of data features (either chemical and/or meteorological) that exhibit discontinuties between the PBL and FT can be used to identify the clusters. Several example profiles will be shown for SH and θ, since these are widely available from radiosonde measurements. In order to remove artifacts due to differences in arbitrary unit scales, each feature is normalized to zero mean and unit standard deviation.

3.2.2 Distances

Multiple methods for defining distances between points provide reasonable results, including: Minkowski 'taxicab' (L^1-norm), Euclidean distance (L^2-norm), and squared Euclidean distance. The L^p-norm distance d_p between two points a and b is calculated from the differences in their m features according to:

$$d_p[a,b] = (|b_1 - a_1|^p + |b_2 - a_2|^p + ... + |b_m - a_m|^p)^{1/p} \tag{3.1}$$

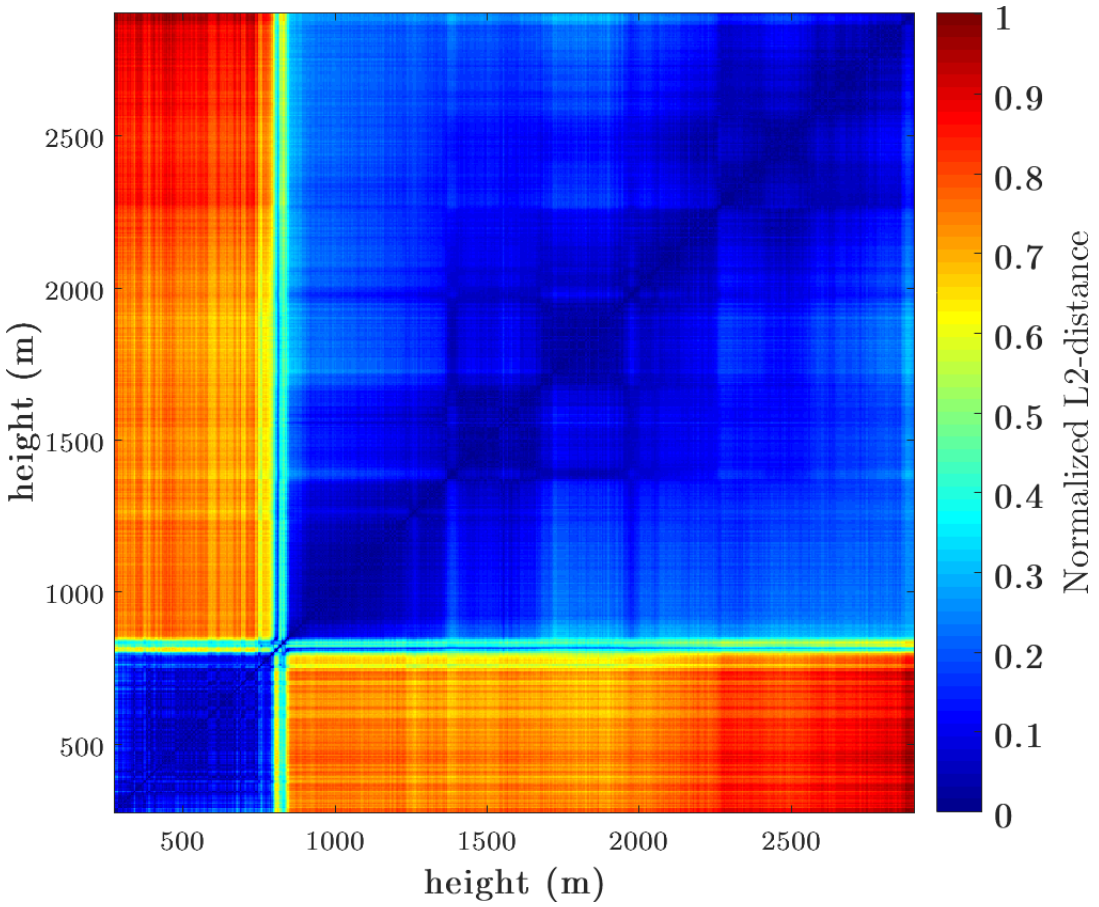

Figure 3.3: L2-norm similarity matrix for a PBL→FT ascent. Clusters appear as self-similar blocks along the diagonal, so z' is visually apparent.

The L^2-norm is used unless otherwise specified.

Figure 3.3 shows the pairwise distance calculated between all points in a single profile, normalized onto (0, 1). The two diagonal blocks of self-similar data points correspond to the PBL and FT layers, whereas inter-layer pairings appear off-diagonal at greater distances.

3.2.3 Forward clustering algorithms

It is possible to start with unlabeled data from a profile and apply various cluster analysis algorithms to search for the optimal partitioning. These methods generally identify the majority of data points correctly, however there are occasional misidentifications (e.g. a few points scattered throughout the FT assigned to the cluster corresponding to the PBL, or

vice versa). The z' can be approximated by finding a height that separates as many points as possible (including erroneous nonphysical assignments). For example, let $\chi_\alpha(h)$ be the fraction of points in α below h, and let $\chi_\beta(h)$ be the fraction of points in β above h. Then the best estimate for height z' is h_{est} such that $\chi_\alpha(h_{est}) \sim \chi_\beta(h_{est})$. Lower values of χ satisfying this equality indicate better clustering assignments.

This approach was used to cluster the data by centroid-based methods (k-means and k-medoids)[2][33][29][44], density-based methods (Density-based spatial clustering of applications with noise "DBSCAN")[19][24], and expectation maximization on Gaussian mixture models.[38]

While all of these models produced reasonable clustering results, the index evaluation methods in §3.2.4 are strongly recommended over these approaches. The forward clustering algorithms described here are sub-optimal in many respects: (1) most are probabilistic and requiring multiple replicates until convergence, (2) they search a larger space of possible subsets that includes many configurations that are not physically possible, so (3) the algorithms tend to find local minima in the space of almost-allowed subsets that must then be used to roughly approximate h_{est} in the space of truly possible configurations.

To the contrary, the index methods presented below are deterministic, so a single sweep across possible partitionings from bottom to top will always provide a single global maximum from the domain of allowed subsets.

3.2.4 Cluster Evaluation Methods

For a vertical profile with N data points, there are slightly fewer than N physically-possible partitionings of the data that could represent the atmospheric layers. A possible partitioning P_n can be generated for each n^{th} data point.

$$P_n \rightarrow \alpha : \{x_1, x_2, ..., x_n\} \ \& \ \beta : \{x_{n+1}, x_{n+2}, ..., x_N\} \tag{3.2}$$

For each of these partitionings, clustering evaluation indices are calculated to measure the 'goodness' of the fit. The P_n resulting in the global maximum for clustering evaluation represents the 'best' partitioning and corresponds to z'.

3.2.4.1 Silhouette criterion values

The silhouette index provides a simple method for cluster consistency evaluation, by calculating how well each point is classified by the assigned memberships.[29][47]

A silhouette value $S(j)$ is calculated for each individual point j by comparing the average distances to other points within the same partition (cluster γ containing n_γ points) and the other partition (cluster φ containing n_φ points).

The average of its distance to points within the same cluster

$$D_\gamma(j) = n_\gamma^{-1} \sum_{x \in \gamma} d_2[x, j] \tag{3.3}$$

and the average distance to points in the other cluser

$$D_\varphi(j) = n_\varphi^{-1} \sum_{x \in \varphi} d_2[x, j] \tag{3.4}$$

are compared and normalized according to

$$S(j) = \frac{D_\varphi(j) - D_\gamma(j)}{\max\{D_\varphi(j), D_\gamma(j)\}} \tag{3.5}$$

to yield a silhouette value between -1 and +1 where higher values indicate a better partitioning.

The overall silhouette index for a given partitioning is determined by averaging the silhouette values of each point. This is typically calculated by assuming empirical prior probabilities and weighting the contribution of each cluster proportionally to its size. However, this results in slight dependency on the ratio of data points measured in the PBL

versus FT. Instead, it is better to use an equal (i.e. non-weighted) average to produce an overall index that is independent of data set boundaries as long as both layers are sampled. While the L^1-norm, L^2-norm, and squared Euclidean distances all provide good results, the use of square Euclidean distance is recommended, since the increased significance of very dissimilar points reduces artificial local maxima near the ends of a profile (though no profiles have been encountered where these small artifacts near the edges result in incorrect identification of the global maximum).

3.2.4.2 Calinski-Harabasz criterion values

The variance ratios within and between clusters are used to calculate the Calinski-Harabasz criterion index for each partitioning.[8] Let m_i be the centroid of cluster i containing n_i data points, and c be the center of the sample data. The within-cluster variance is calculated according to

$$D_W = \sum_{x \in \alpha} (d_2[x, m_\alpha])^2 + \sum_{x \in \beta} (d_2[x, m_\beta])^2 \tag{3.6}$$

and the between-cluster variance by

$$D_B = n_\alpha (d_2[m_\alpha, c])^2 + n_\beta (d_2[m_\beta, c])^2. \tag{3.7}$$

The overall Calinski-Harabasz index is calculated as:

$$CH = (n_\alpha + n_\beta - 2) D_B D_W^{-1}. \tag{3.8}$$

3.3 Results

3.3.1 Forward Clustering

For brief illustration, figure 3.4 shows the application of k-medoids clustering on the first ascent from the SONGNEX 2015.04.13 flight over Denver, Colorado, USA. The right

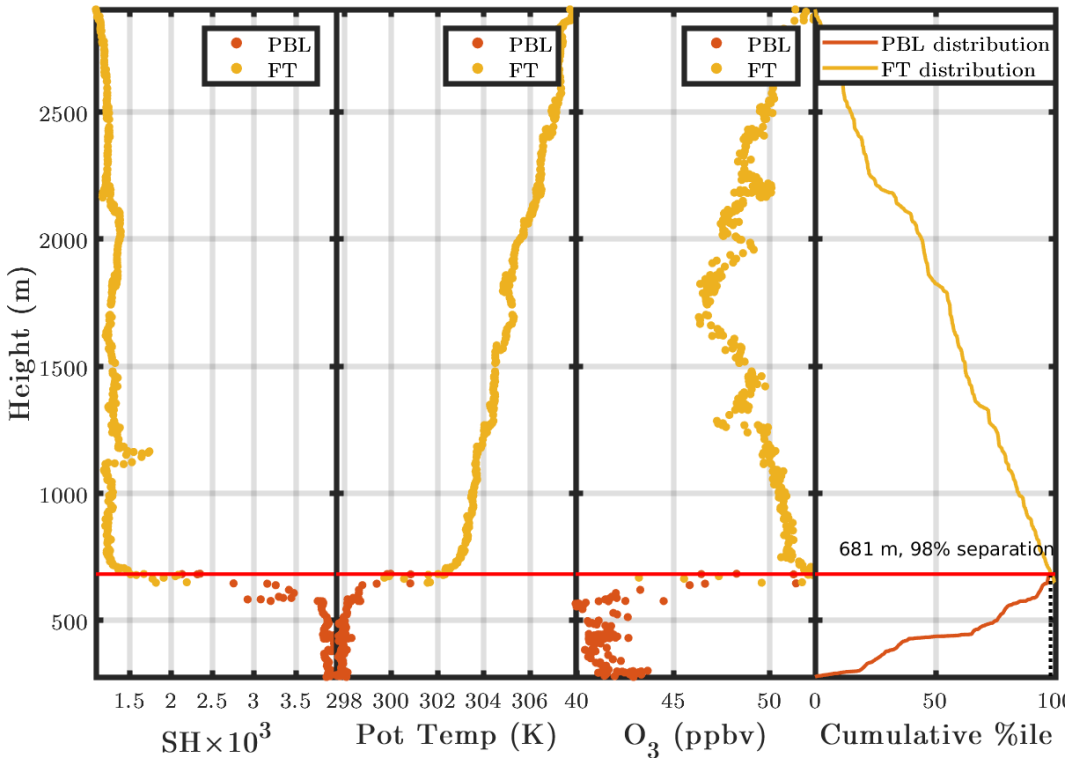

Figure 3.4: Specific humidity (SH), potential temperature (θ), and ozone (O_3) for the first ascent on the SONGNEX 2015.04.14 flight. The right pane shows orange $\chi_\alpha(h)$ and yellow $\chi_\beta(h)$ traces, along with red line indicating h_{est}

panel shows $\chi_\alpha(h)$ and yellow $\chi_\beta(h)$ traces, which cross at h_{est} = 681 m, where $\chi_\alpha(h_{est}) \sim \chi_\beta(h_{est}) \sim 0.02$.

3.3.2 Comparison of Indices

Figure 3.5 demonstrates the application of silhouette values and the Calinski-Harabasz index to the first three ascents during the SONGNEX flight (top to bottom). Both cluster evaluation measures identify ideal partitionings corresponding to z'. The greatest difference between methods is a negligible 35 m uncertainty on the first profile. Moving forward, the Calinski-Harabasz index is used to select the best partitioning. Note extreme variability in SH for the third ascent; the wide range of values in the PBL may confuse a forward clustering algorithm. However the resulting irregularities in the evaluation indices occur away from the global maximum and consequently have no impact on the determination of

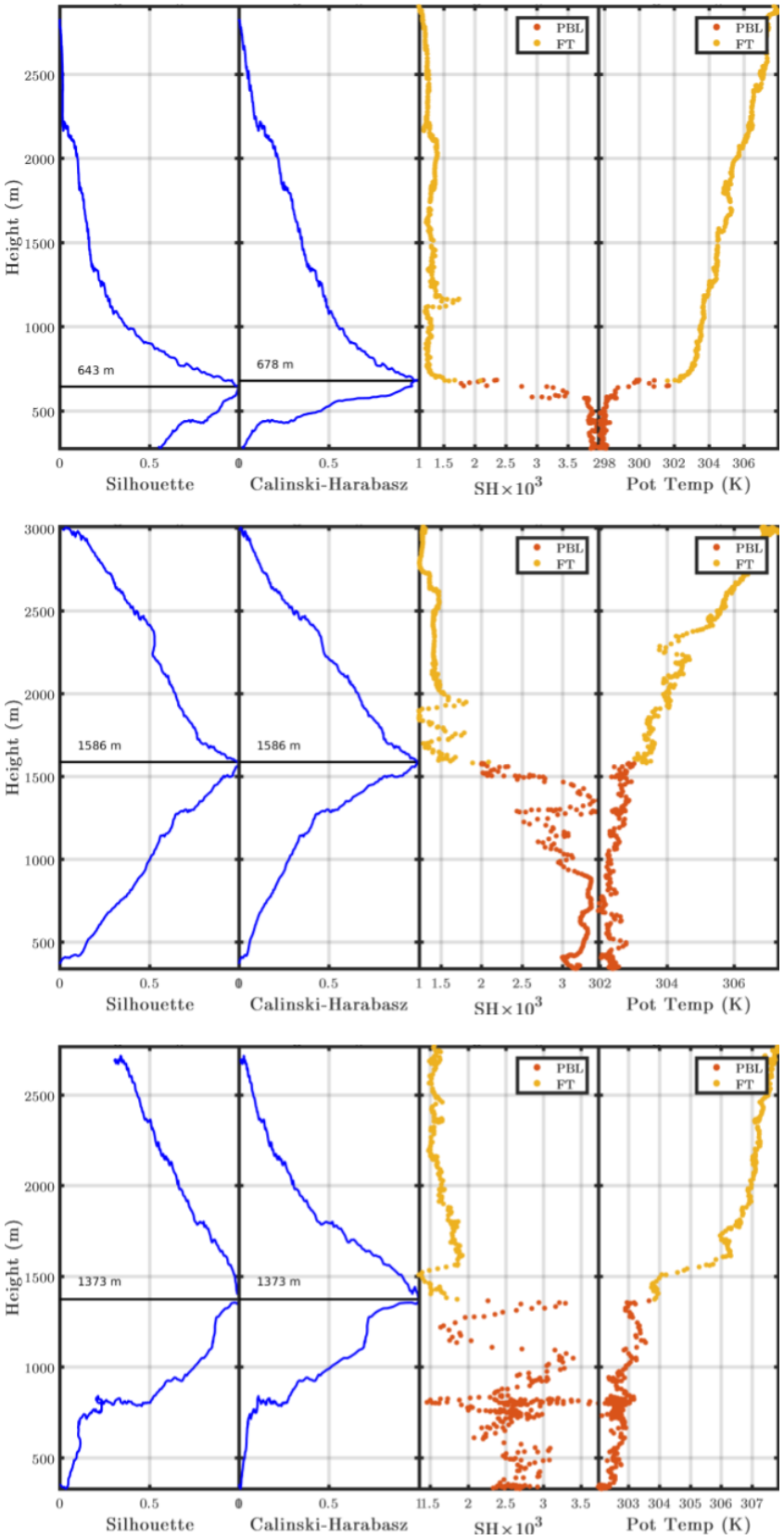

Figure 3.5: Specific humidity (SH) and potential temperature (θ) for the first three ascents on the SONGNEX flight. The silhouette values and the Calinski-Harabasz indices for each partitioning are shown at left, along with the height corresponding to their global maximum.

z'.

Figure 3.6 shows that identification of two-cluster topology is relatively independent of selection of feature subsets, as long as they are appropriately decoupled between layers. The top panel shows z' calculated as in figure 3.5. The middle panel shows identification using only trace gases (ozone, formaldehyde, and carbon monoxide). The bottom panel shows application to only volatile organic compounds measured by chemical ionization mass spectrometry. These calculations using mutually-exclusive subsets of the data all agree $z' = 665 \pm 25$ m.

Figure 3.7 shows application to the first three ascents of the 2013.05.19 flight during Pan-European Gas Aerosol Climate Interaction Study (PEGASOS) over Jämijärvi, Finland. The relative humidity (RH), potential temperature (θ), and hydroxyl reactivity measurements (kOH) are selected for evaluation. Even with sparse kOH measurements relative to the 1 Hz meteorological measurements, z' is easily identified evaluating the ~ 35 all-present data points.

3.4 Conclusions

Methods from cluster analysis are adept at identifying the boundary in the two-cluster topology that arises natural from data sets that include measurements from both the PBL and FT. The silhouette value and Calinski-Harabasz indices evaluated over all physically-possible subsets produce a global maximum at the altitude corresponding to the height of the planetary boundary layer (z'). These methods are robust to measurement noise, feature selection (see figure 3.6), and minor anomalies in atmospheric structure (see SH in bottom panel of figure 3.5).

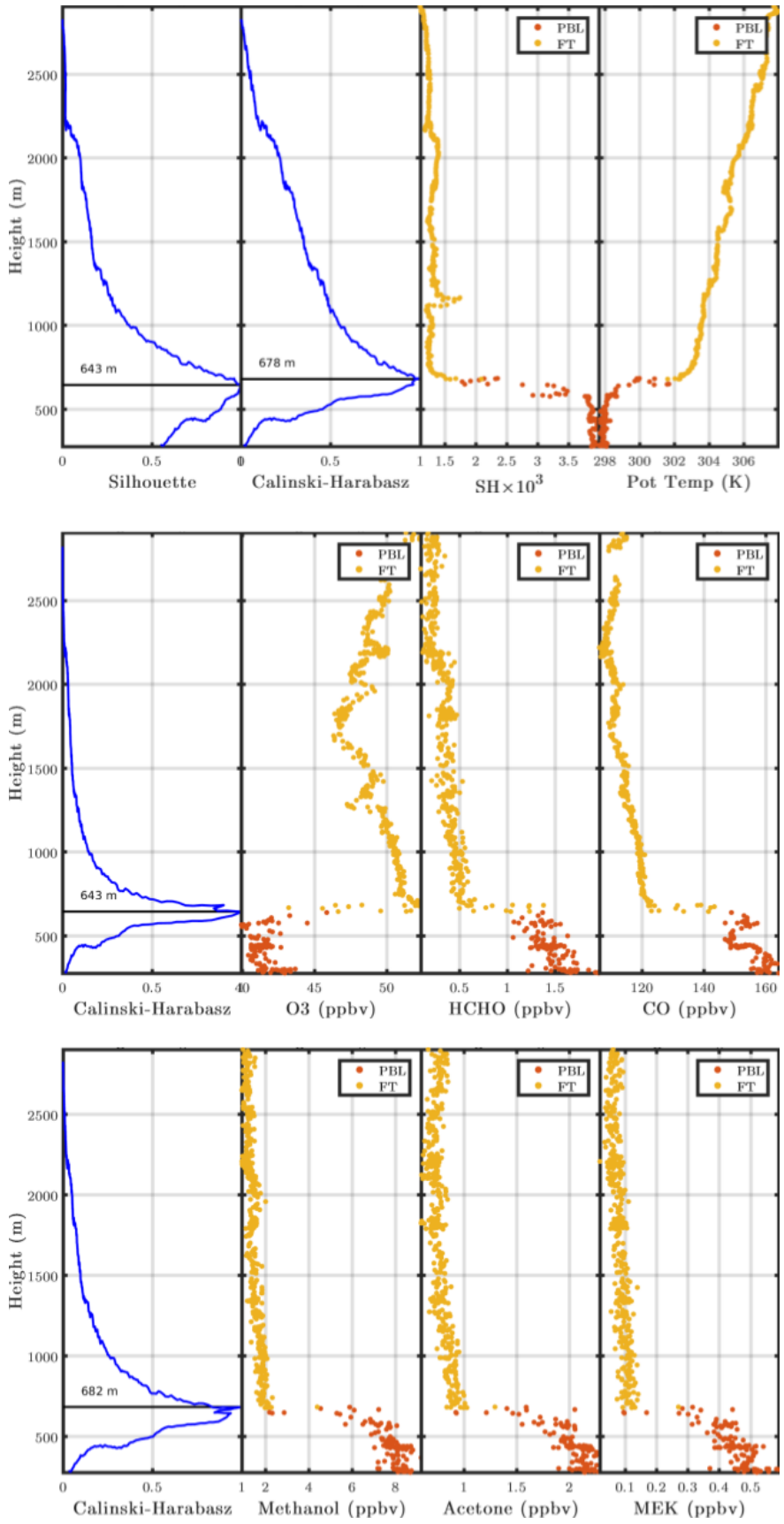

Figure 3.6: The Calinski-Harabasz index method is used to find z' for the first SONGNEX ascent using three mutually-exclusive subsets of measurement types: meteorological (top), trace gases (middle), volatile organic compounds (bottom).

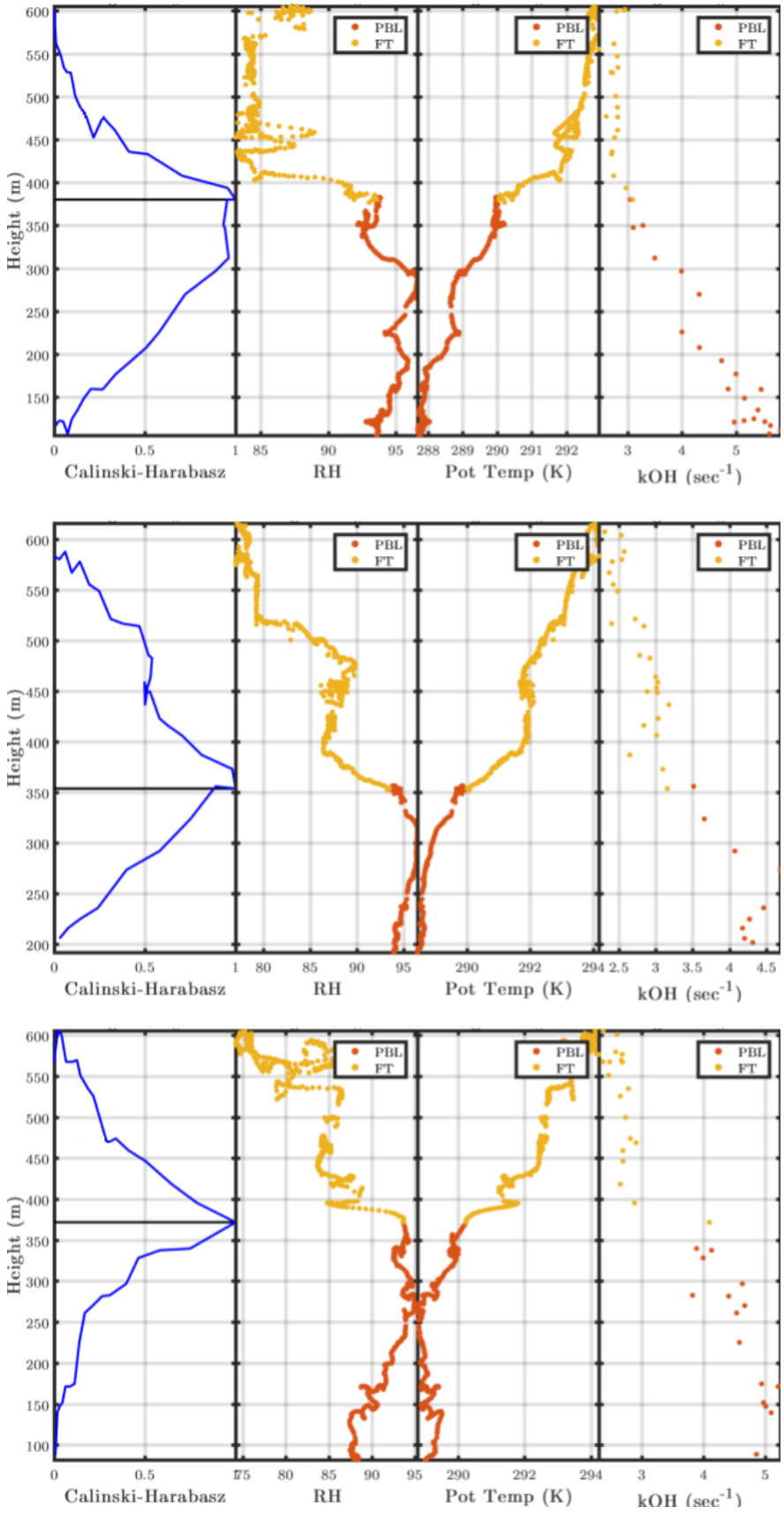

Figure 3.7: Relative humidity (RH), potential temperature, and hydroxyl reactivity measurements (kOH) during the first three ascents on the PEGASOS flight are evaluated by the Calinski-Harabasz value to find z'.

3.5 Future Work

A natural next step is investigation of atmospheric structure above PBL/FT boundary. If other layers are sufficiently decoupled, the principle should hold and produce reasonable results.

The 'In-service Aircraft for a Global Observing System' (IAGOS) project equips commercial aircraft with instruments to measure several trace gases (O_3, CO, CO_2, CH_4, NOx, NOy, H_2O) in addition to the meteorological data necessary for navigation. These flights produce vertical profiles from airports around the world on an hourly or daily basis. This data set could be assimilated with the methods described here to generate continuously-updating estimates of planetary boundary height near airports around the world.

4 ASSESSMENT OF SATELLITE CAPABILITIES FOR DISCERNING HCHO, O_3, AND NO_2 ENHANCEMENTS FROM MULTIPLE SOURCES

ABSTRACT: Instruments to be deployed on new and upcoming satellites, such as TROPOMI (2017 launch) and TEMPO (2018/19 projected launch), will be capable of measuring air quality-relevant species at unprecedented spatial and temporal scales. However, studies on the degree of detail that this will afford are lacking. This work addresses the question: Given a region with a complex scene featuring signatures from multiple types of anthropogenic activity, what spatial resolution and chemical precision are necessary to resolve plumes and distinguish between emissions types? These specifications are studied for formaldehyde (HCHO), nitrogen dioxide (NO_2), and tropospheric O_3.

Chemical and physical parameters measured via aircraft in the boundary layer and free troposphere (FT) during the Shale Oil and Natural Gas Nexus (SONGNEX 2015) field campaign are employed to view chemical enhancements from biomass burning and urban outflow over the region northeast of Denver, Colorado. The spatially and temporally resolved in-situ data are used to calculate the planetary boundary layer contributions to the column densities for HCHO, O_3, and NO_2. The converted data are mapped with varying constraints imposed to mimic measurement limitations. Pixel footprint resolution is probed using 2D spatial bins, and the loss of detail due to uncertainty in measurements is emulated by 1D signal bins to mask gradients that are finer than the instrumental precision.

First, the spatial resolution and chemical precision limits are studied for each species. Second, the scene is emulated using the specifications for TROPOMI and expected performance for TEMPO, in order to assess the degree to which their retrievals will be able to discern the signatures of various activities, and to ascertain the information that may be derived from trace gas enhancements.

4.1 Introduction

Satellites provide a wealth of information pertinent to atmospheric chemistry and air quality, including both meteorological and chemical data. The types of scientific questions that can be answered with a given data set depend on its spatial and temporal coverage, and the quality of the measurements.[32][58] With each new generation of remote-sensing instruments, the level of detail recorded improves dramatically, both in terms of decreasing pixel footprint and increasing sensitivity.[4][7][54][66]

This work considers satellite measurements of formaldehyde (HCHO), ozone (O_3), and nitrogen dioxide (NO_2). Tropospheric O_3 is an EPA criteria pollutant that is detrimental to both agricultural activities and human respiratory health.[1][37][55] Formaldehyde is a valuable tracer for multiple emissions types and photochemical processes related to the production of tropospheric O_3 and secondary organic aerosol.[11][25][28][61][13] In the biomass burning plume studied below, HCHO is strongly correlated ($r \sim 0.9$) with fine particulate matter, another EPA criteria pollutant that damages both the respiratory and cardiovascular systems.[15][52][17]

Spatial resolution is an important limit to the utility of satellite data since pollutants vary on scales smaller than a satellite pixel and impact regions' chemistry and air quality in ways not apparent from the measurements. Likewise, the instrumental precision of the chemical measurements plays a large role in determining which complex features of air chemistry can be detected.[20]

These limits are investigated using data collected during a Shale Oil and Natural Gas Nexus (SONGNEX 2015) field campaign research flight that systematically mapped the spatial distribution of pollutants in the planetary boundary layer over a 80 x 70 km region. The flight track pattern provided ideal conditions for converting in-situ measurements to column density contributions in order to systematically regroup data into various configurations that emulate instrumental limitations for remote-sensing platforms.

4.2 Methods

4.2.1 Measurements

Flights for the SONGNEX campaign were conducted during April-May of 2015 onboard the NOAA WP-3D research aircraft. Details regarding the measurements, platform, flight plans, and research goals are available elsewhere.[42]

Measurements for HCHO were obtained using the NASA In Situ Airborne Formaldehyde (ISAF) instrument with 10% accuracy.[9] The O_3 data were obtained by NO-induced chemiluminescence with 40 pptv + 5% uncertainty.[49][46] Measurements for NO_2 were obtained using the NOxCaRD cavity ringdown absorption spectrometer with $\pm$5% accuracy.[21] Measurements for CH_4 were obtained by IR laser absorption in a high-finesse cavity (Picarro) with $\pm$1.2âŁ‰ppb accuracy.[45][12] A nucleation-mode aerosol size spectrometer (NMASS) and ultra-high sensitivity aerosol spectrometer (UHSAS) measured several properties of fine particulate matter with optical diameter 0.004 - 1.0 μm: number concentration ($\pm$9% accuracy), surface area (+22%/-12% accuracy), and volume (+36%/-18% accuracy).[6][5] All measurements were collected at 1 Hz.

Occasional gaps (<50 s once per hour) in observed HCHO were estimated by regression using well-correlated tracers: acetaldehyde ($r = 0.84$), acetic acid ($r = 0.90$), furfural ($r = 0.86$) and propionic acid ($r = 0.84$). These were measured by H_3O^+ Time-of-Flight Chemical Ionization Mass Spectrometry (H_3O^+-ToF-CIMS).[30][62]

4.2.2 Analysis

4.2.2.1 Converting mixing ratio to column density

While the research flights measure in-situ mixing ratio (χ, e.g. parts per billion), a satellite measurement is integrated along its line of sight with results reported as vertical columns (D, e.g. molecules/cm^2). To calculate a species's contribution from the planetary boundary

layer to the measured vertical column (D_{PBL}), the number density (n, e.g. molecules/cm^3) is integrated from the ground ($z = 0$) to the top of the PBL ($z = z'$).

$$D_{PBL} = \int_0^{z'} dz\, n(z) \tag{4.1}$$

Assuming a well-mixed PBL, the vertical profile for number density is calculated from measured mixing ratios by scaling vertical pressure according to the hydrostatic equation. Thus,

$$D_{PBL} = \frac{N_A P_\circ \chi}{RT} \int_0^{z'} dz\, e^{-z/h} \tag{4.2}$$

$$= \frac{h N_A P_\circ \chi}{RT} (1 - e^{-z'/h}) \tag{4.3}$$

where $P_\circ$ is the pressure at $z = 0$, T is temperature, h is the scale height of the atmosphere, N_A is Avagadro's number, and R is the gas constant.

The height of the PBL (z') was calculated from collocated gradients in multiple measurements during periodic spirals from approximately 500 - 2000 m over fixed points. The transition from the PBL to the FT is clearly delineated by a sharp decrease in specific humidity (SH) accompanied by increased potential temperatures (θ), which correspond with gradients in HCHO and O_3. Figure 4.1 shows an example of the measurements during the first spiral ascent, with the top of the PBL indicated by the red line.

4.2.2.2 Satellite Emulation

While the aircraft measurements provided chemical information on the scale of hundereds of meters, satellite pixel footprints are on the scale of multiple kilometers. In following plots that emulate data loss to spatial resolution limitations, the data collected in the PBL are grouped into 2D bins representing satellite pixels. Analogously, the research flight's in-situ instruments measured very fine chemical gradients, whereas satellites have a much coarser precision. In order to represent how chemical precision limits mask fine variability, the data

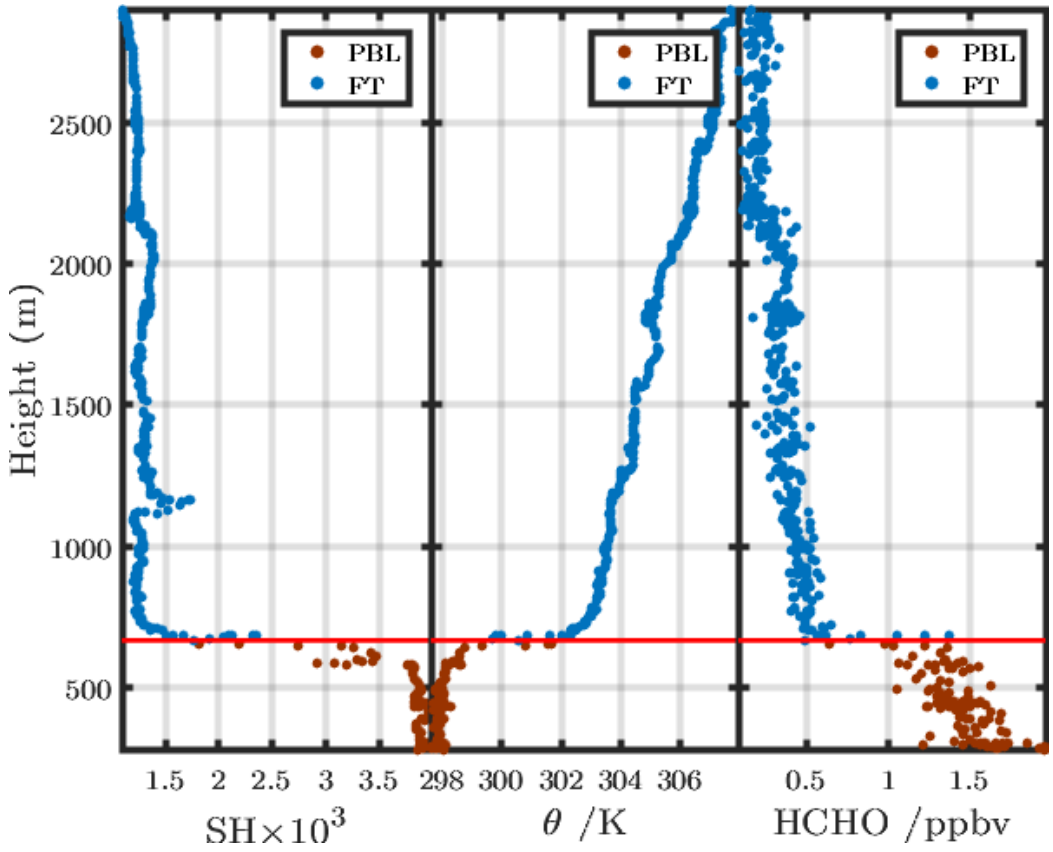

Figure 4.1: The height of the PBL during a given profile is easily identified from discontinuities in specific humidity (SH) and potential temperature (θ). These measurements from the first PBL→FT ascent are shown here as a function of height above ground level, alongside HCHO for comparison.

are grouped into bins by concentration. In the same way that two data points separated by 50 m are highly unlikely to be resolved in a 2 km satellite footprint, two measurements differing in value by 0.5% are highly unlikely to be meaningfully resolved on an instrument with 15% precision. These methods for emulating measurement constraints are used to answer two different questions: (i) what spatial resolution and chemical precision are required to capture various features of the anthropogenic activity in the PBL, and (ii) what kind of information are likely to be resolved from new and upcoming satellites, specifically TROPOMI and TEMPO.

To answer the latter question, those particular satellites' instruments were emulated by binning the data according to the specifications for TROPOMI [54][18][57] and expected

Table 4.1: Emulations of satellite instruments are based off of TROPOMI specifications and TEMPO expected performance.

	TROPOMI	TEMPO
Spatial Resolution	7x7 km	2.1x4.7 km
Revisit frequency	Daily	Hourly
HCHO precision	$1x10^{16}cm^{-2}$	$0.4x10^{16}cm^{-2}$
PBL O_3 precision	25%	2.6 ppb

performance of TEMPO.[66] Satellite specifications are collected in Table 4.1. TROPOMI is sun-synchronous and provides global daily coverage with an early afternoon overpass. TEMPO is geostationary and scans across Greater North America, providing hourly pollutant measurements (the hourly HCHO, SO_2, and CHOCHO retrievals are averaged and reported 3 times/day).

4.3 Results and Discussion

4.3.1 Flight and Observations

This assessment is based on data collected on 2015.04.13 north of Denver, shown in figure 4.2 (colored by HCHO mixing ratio). The P-3 flew a raster pattern in the PBL over an 80 km (E-W) x 70 km (N-S) region, with 5 km N-S spacing and four spirals into the FT over fixed points at the edges of this work's field of regard.

The flight sampled three distinct chemical regimes, which are geofenced and designated: 'Biomass Burning' [40.485 - 40.635 N, 104.510 - 104.910 E]; 'Urban Outflow' [40.020 - 40.1450 N, 104.541 - 104.775 E]; and 'Background', [40.270 - 40.611 N, 104.250 - 104.346 E]. Figure 4.3 shows the difference between regions for HCHO, O_3, NO_2, and CH_4. Both polluted regions contained HCHO enhancements up to 4 ppb. The biomass burning region contained elevated mixing ratios of ozone. The 'urban outflow' plume cointained a mixture of anthropogenic pollution sources: emisson from Denver (note enhanced NO_2) mixed with

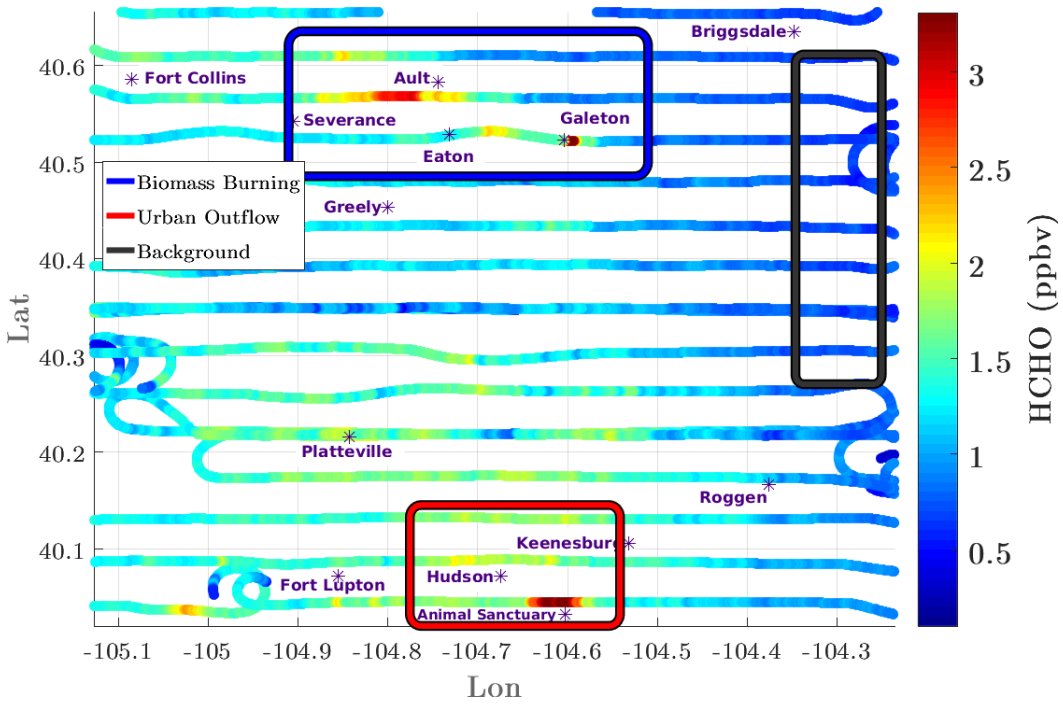

Figure 4.2: Top-down view of the 2015.04.13 flight track, cropped to the portion of the raster pattern that will be used for satellite emulation. The flight track is colored by formaldehyde, and the three regions of interest are indicated.

some burning emissions and recirculation of air affected by oil and natural gas activity in the Denver-Julesburg basin (note enhanced CH_4). This region included less ozone relative to the other regions, however this is also partially due to the fact that this region was sampled first, leaving less time for photochemical formation.

The burning emissions were due to a fire on farmland 1 km south (upwind) of Galeton, CO. The source of the plume is captured by the video camera on the P-3, and that event is the only visible burning in that region of the flight. As shown in figure 4.4, HCHO in the burning plume correlated well ($r \sim 0.9$) with volume of fine particulate matter (0.004 to 1.0 μm diameter), indicating potential for localizing particulate matter pollution events using HCHO as a tracer. The plume was detected on four legs of the flight spanning >30 min, after which the P-3 left the area.

As shown by the flight track in figure 4.5, most of the flight mapped the spatial distribution of species in the PBL. The plane spiraled into the FT four times, spaced out

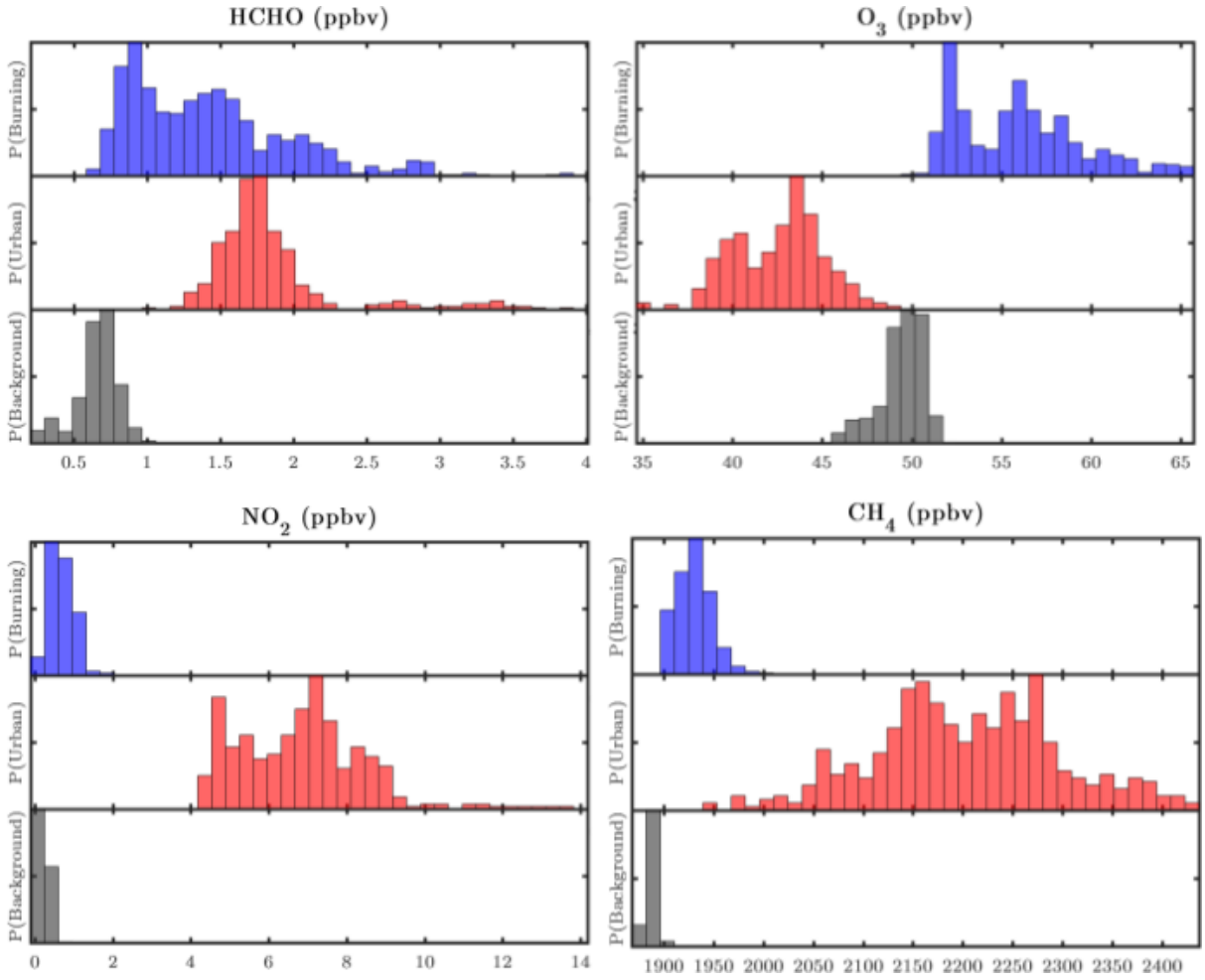

Figure 4.3: Histograms of ambient concentrations in the biomass burning (blue), urban outflow (red), and background (gray) regions, sorted by chemical starting clockwise from upper left: HCHO, O_3, CH_4, NO_2.

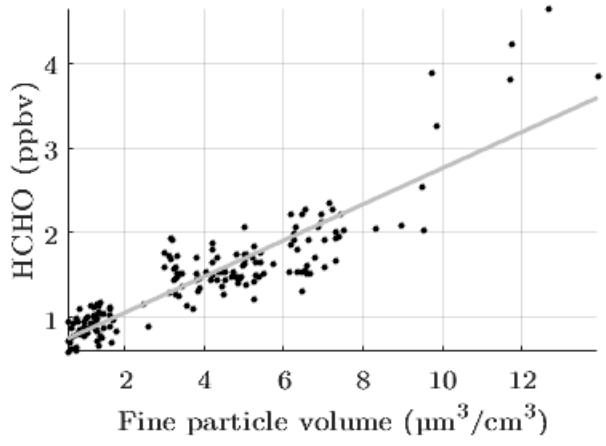

Figure 4.4: HCHO is a tracer for particulate matter in the biomass burning plume

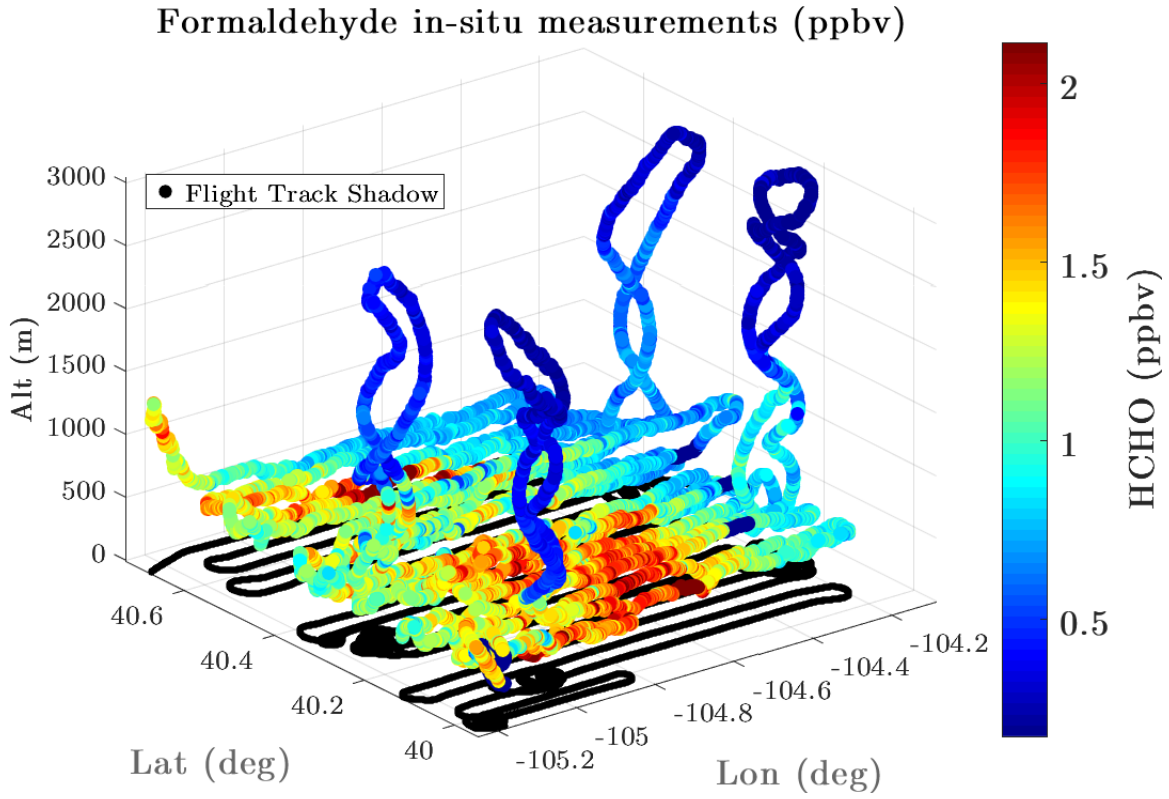

Figure 4.5: A 3-dimensional view of the flight track, showing the spirals at each corner. high formaldehyde concentrations and variability are limited to the planetary boundary layer.

geographically and throughout the flight. The ascent and descent over a fixed point provided a full vertical profile, and allowed determination of the PBL height. As shown by the flight track color, HCHO in the PBL exhibits significant spatial variability. However the FT contains lower concentrations even over regions with enhancements in the PBL, suggesting that the layers are relatively decoupled. Figure 4.6 shows the vertical distribution of HCHO over the course of the entire flight, highlighting the greater spatial variability and concentrations in the PBL relative to the FT, which was relatively spatially and temporally homogenous. Thus the FT simply adds a background offset to satellite measurements, and pollutants in the PBL will drive spatial variability in the column. Therefore this assessment of satellite capability to resolve various activities will focus on their PBL contributions to total column.

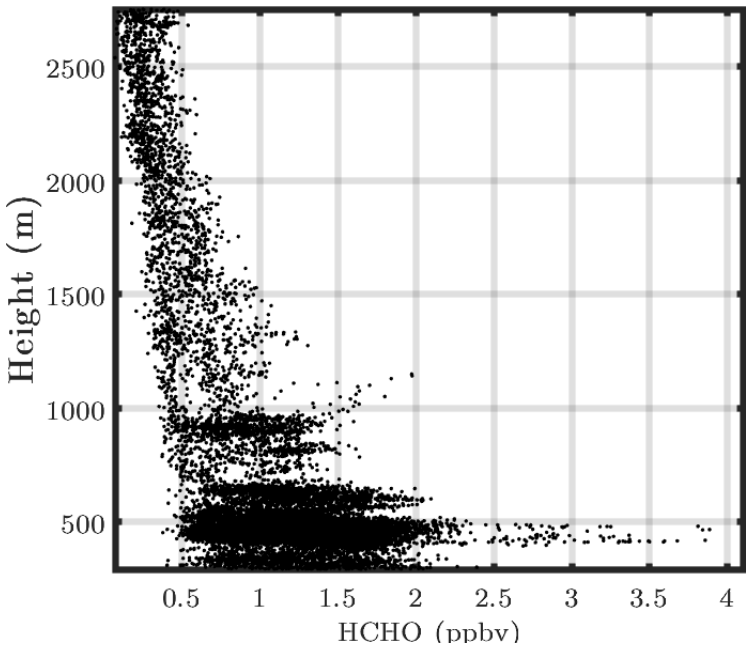

Figure 4.6: Vertical distribution of HCHO over the entire flight. High concentrations and spatial variability are seen in the PBL, while the FT has lower concentrations and is relatively homogenous throughout the flight.

4.3.2 Spatial Resolution

Spatial resolution limits the scale of features that can be resolved from space; consequently the ever-decreasing pixel footprints of newer instruments are dramatically increasing the level of detail available. For each species, the flight measurements were used to emulate satellite retrievals while varying pixel footprint sizes over two orders of magnitude to study at which resolutions various features become discernible. Several levels of detail for HCHO are included for illustration, along with final results for other species.

Figure 4.7 shows the high-resolution 'satellite-view' of HCHO variability (i.e. converted to column density from the 1-Hz resolution ($\sim$80 m) and 10 pptv accuracy data). With the highly-resolved data, there is a strong HCHO signature from the biomass burning (ΔHCHO $\sim$1x10^{16}cm^{-2}), and a field of minor HCHO enhancement (ΔHCHO $\sim$5x10^{15}cm^{-2}) around the Denver region. The HCHO hotspot over Denver appears less prominent relative to figure 4.2 since the high concentration is mixed into a shallow boundary layer ($\sim$650

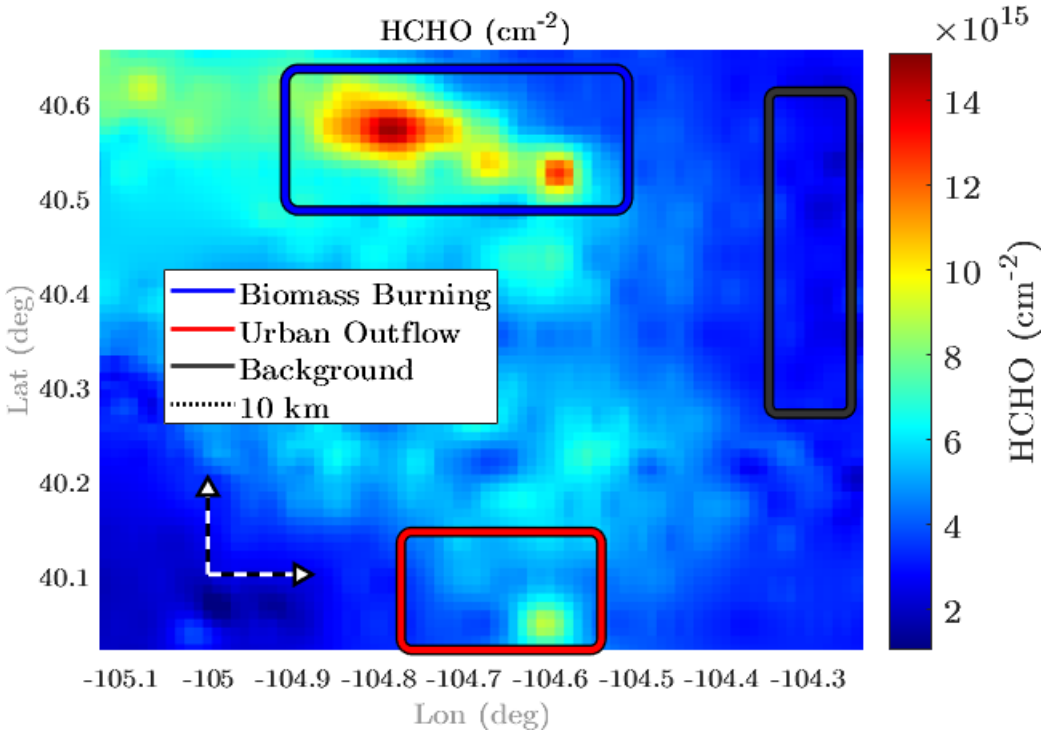

Figure 4.7: High-resolution (1x1 km) distribution of HCHO in the PBL in the region of interest. This shows the emulated satellite perspective with no imposed constraints limiting spatial resolution or chemical precision.

m). For the remainder of §4.3.2, these data will be spatially binned to varying pixel sizes to ascertain the level of resolution necessary to discern these features.

Figure 4.8 shows the HCHO distribution at varying resolutions. As shown in the upper left pane, a minimum resolution of 20 x 20 km is necessary to discern the enhancements near the fire and urban plume. With this coarse view, the range of concentrations measured is greatly attenuated, since the signals from localized hotspots appear diluted by nearby measurements. As indicated by the arrows on the colorbar, the magnitude of the hotspot is underestimated by a factor of two.

At 10 x 10 km (upper right) the burning plume and field of enhancement north of Denver are discernible, and regional trends are clear. The intensity of the enhancement from biomass burning is still attenuated, but less dramatically ($\sim$10x10^{15}cm^{-2}, compared to the true 15x10^{15}cm^{-2}). Decreasing the pixel footprint to 5 x 5 km (lower left), a relatively

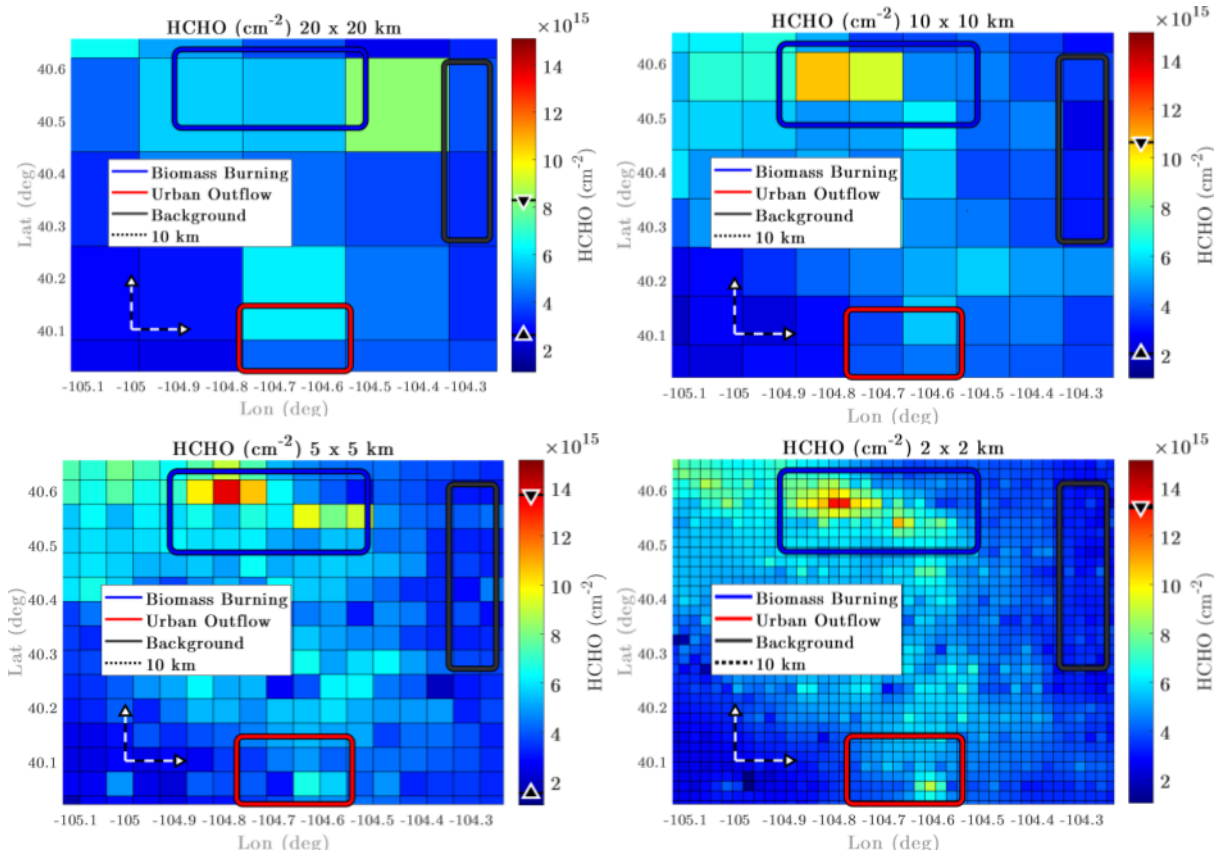

Figure 4.8: The effect of spatial resolution on the ability to discern chemical enhancements, shown for HCHO binned to 4 pixel footprints (clockwise starting from upper left: 20x20km, 10x10km, 5x5km, 2x2km). The color scale is the same in each figure, and the arrows on the colorbar indicate attenuation of the observed concentration ranges as variations are averaged into larger pixels, losing information about the extremes.

nuanced image emerges, detailing both plumes and capturing the intensity of the hotspot. At 2 x 2 km (lower right) the satellite would capture the same level of detail as the high-resolution view. 2 x 2 km is the lower bound for pixel sizes that can be probed with this data set, since this method is limited by the distance between legs of the flight.

This analysis was completed for several satellite-viable species; for all studied, a pixel footprint size on the order of 5 x 5 km was found to capture sufficiently detailed information about spatial variability and most of the chemical range. This resolution captures significant detail in the PBL O_3, showing the depletion in the Denver plume and enhancement from the burning plume clearly against the background (figure 4.9). The NO_2 is also well resolved

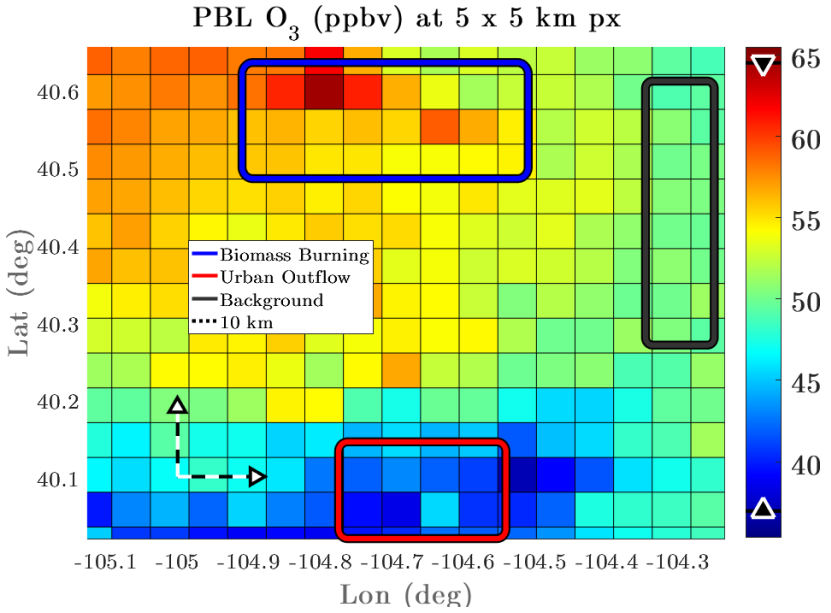

Figure 4.9: Tropospheric O_3 at 5x5 km resolution

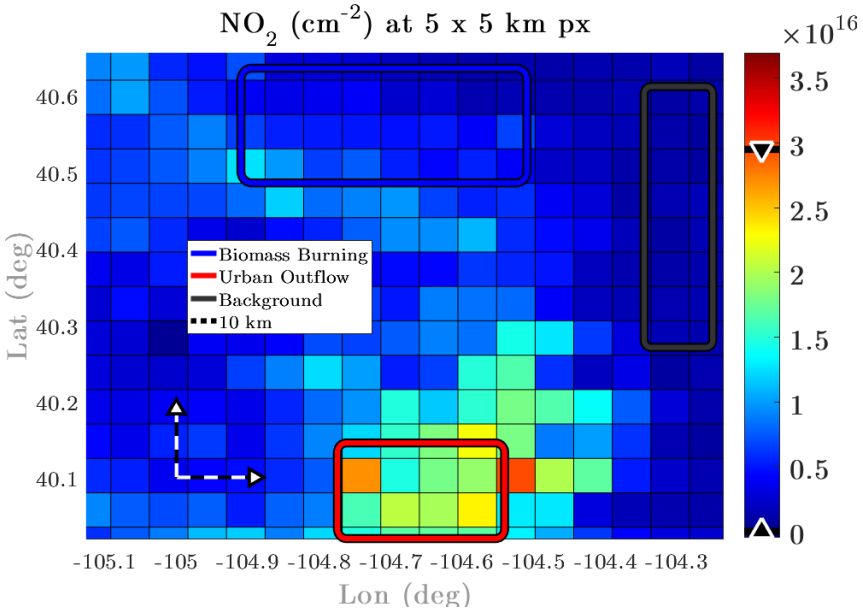

Figure 4.10: Tropospheric NO_2 at 5x5 km resolution

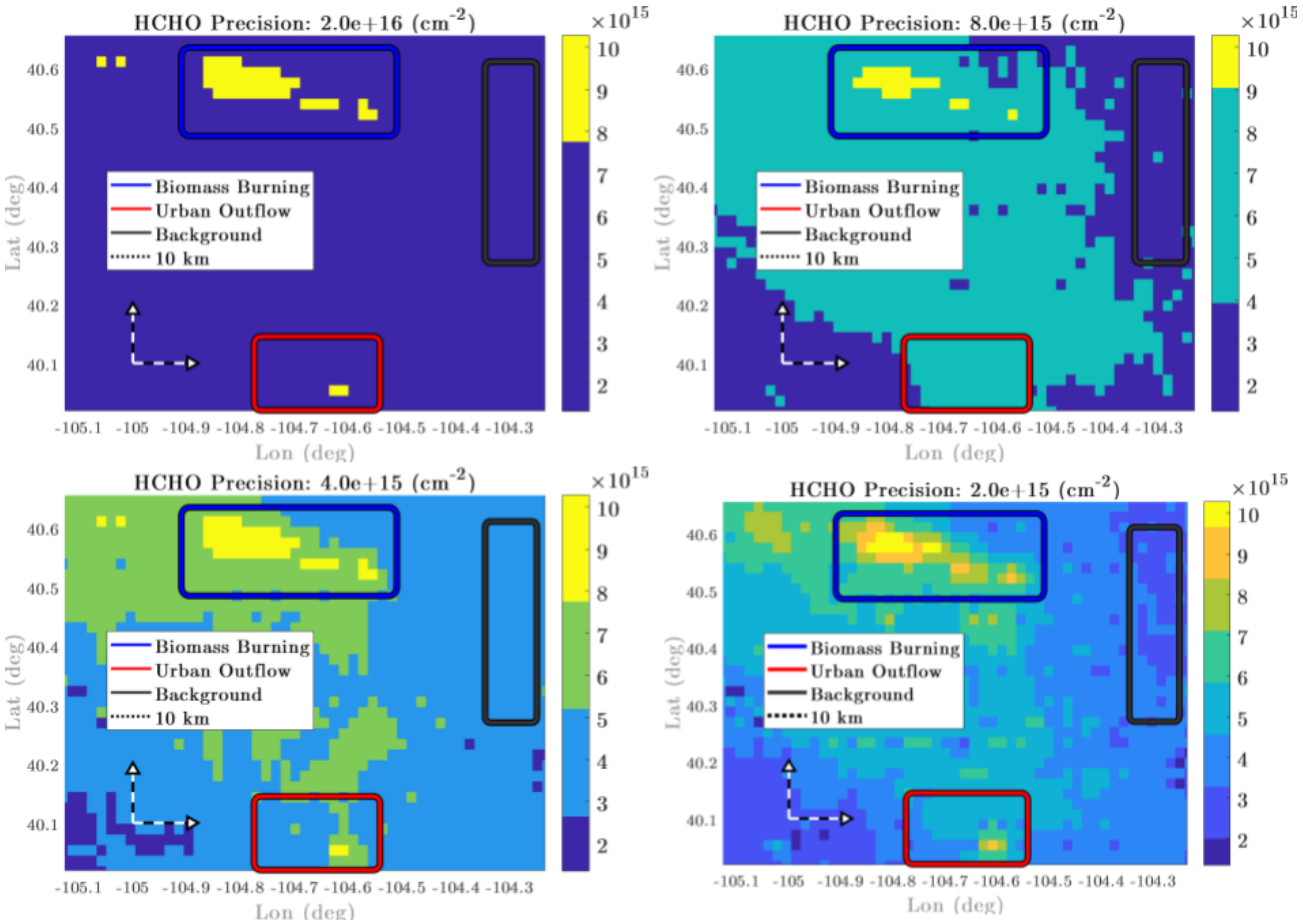

Figure 4.11: The effect of chemical precision (signal uncertainty) on the ability to discern chemical enhancements, shown for HCHO binned from most coarse to most precise (clockwise starting from upper left: $2x10^{16}cm^{-2}$, $8x10^{15}cm^{-2}$, $4x10^{15}cm^{-2}$, $2x10^{15}cm^{-2}$)

at 5 x 5 km, showing details of the plume shape and distribution (figure 4.10).

4.3.3 Chemical Precision

Spatial resolution is only one aspect to evaluate for assessing satellite capabilities; it is necessary to also consider the measurements' chemical uncertainty. Limits due to chemical precision were studied in a manner analogous to spatial resolution. In §4.3.2 high resolution (e.g. 1 m) data was averaged into larger (e.g. 20 km) bins, to emulate how (for example) a 50 x 50 m feature isn't visible in a 5 x 5 km pixel. In §4.3.3 high precision (e.g. pptv) data are averaged into larger (e.g. ppbv) bins. This is meant to emulate how a 1 ppbv gradient

Table 4.2: Precision necessary to a) discern broad plumes and b) resolve finer details

Resolve:	Plumes only	Fine variability
HCHO	2x10^{16}cm^{-2}	4x10^{15}cm^{-2}
PBL O_3	15 ppb	5 ppb
NO_2	1x10^{18}cm^{-2}	5x10^{17}cm^{-2}
CH_4	2x10^{16}cm^{-2}	1x10^{15}cm^{-2}

will not be visible to a satellite with 10 ppbv precision.

Figure 4.11 shows this type of analysis applied to the HCHO measurements. While 2x10^{16}cm^{-2} precision (upper left) registers the burning plume, the minimum precision necessary to resolve the basics of the scene is 8x10^{15}cm^{-2} (upper right), at which point the general enhancements relative to the background register. At 4x10^{15}cm^{-2} (lower left) precision it becomes possible to distinguish the finer structure of the two plumes, and between them an enhancement relative to the background region. Most of the HCHO features measured in-situ can be captured well with precision as coarse as 150 pptv, which corresponds to a difference in column of approximately 2x10^{15}cm^{-2} (lower right). Consequently, at this precision it would be possible to resolve most key details of the chemical gradients observed during the flight. Analogous analysis on other satellite-relevant species identified the precision necessary to a) discern broad plumes and b) resolve finer details, with results in table 4.2.

4.4 Satellite Comparisons

To place these results in the context of actual satellite platforms, various measurements were emulated for upcoming instruments, taking into account their expected spatial resolution and chemical precision to assess whether contributions to columns from the PBL will be discernible. The expected level of detail from TEMPO and TROPOMI is visualized in figure 4.12.

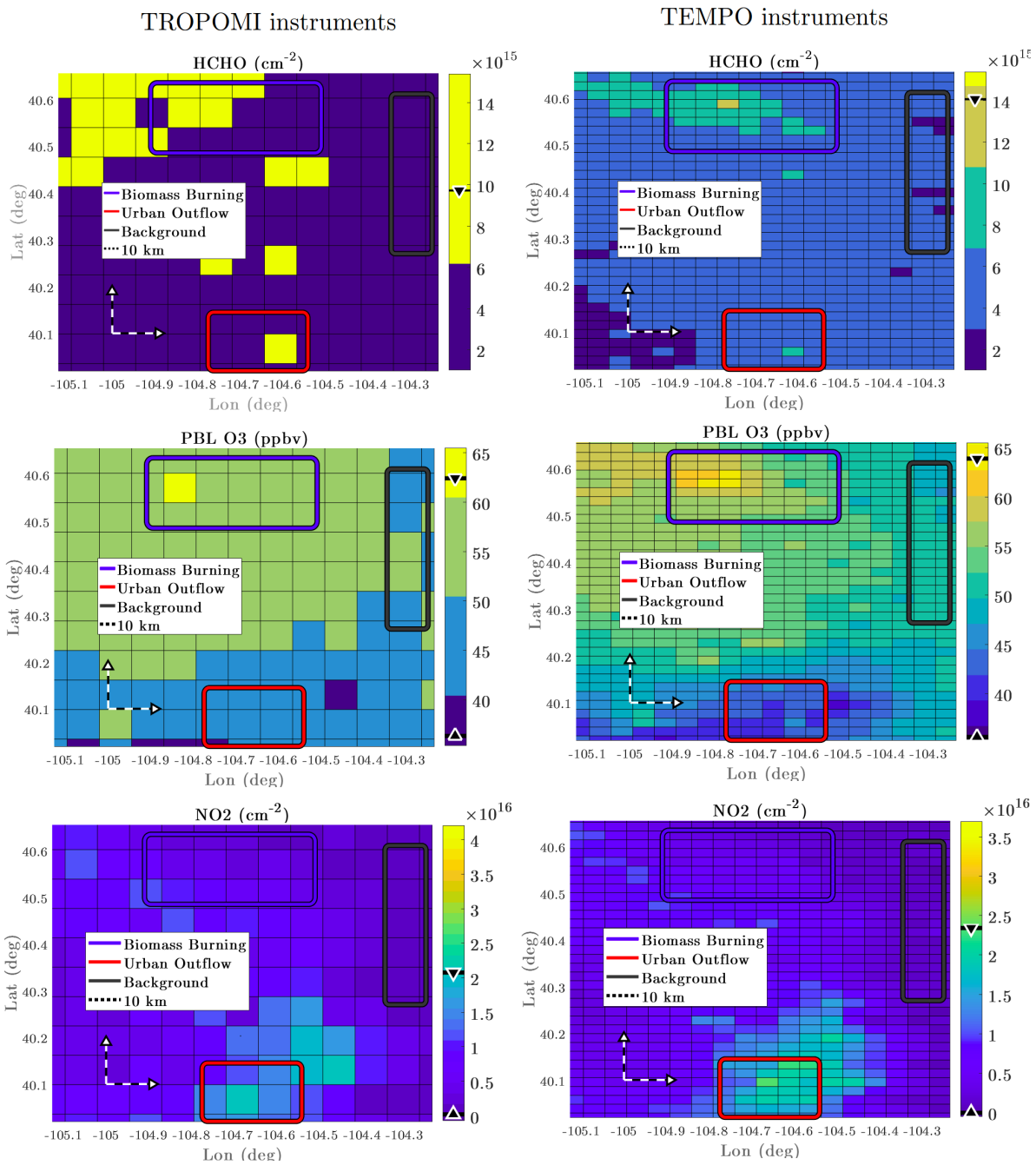

Figure 4.12: Expected TROPOMI (left) and TEMPO (right) perspectives on the region of regard, emulating limitations described in Table 4.1. From top to bottom, HCHO, O_3, and NO_2. The arrows on the colorbars indicate attenuation of the observed concentration ranges as variations are averaged into larger pixels.

4.4.1 OMI

The Ozone Monitoring Instrument (OMI) currently delivers daily measurements from aboard the NASA Aura platform. Because the spatial resolution is on the same order of magnitude as the field of regard for this work, the ozone distribution registers as a few pixels with the same average value. The actual O_3 measurements from OMI on 2015.04.13 were retrieved via the NASA EarthData interface. All pixels in the region of interest registered $\sim$325 DU with no spatial variability.

4.4.2 TROPOMI

The TROPOspheric Monitoring Instrument (TROPOMI) is a Dutch satellite instrument launched on the Copernicus Sentinel-5 Precursor satellite in 2017. TROPOMI will measure O_3, CH_4, HCHO, aerosol, CO, NO_2 and SO_2 with 7 x 7 km daily global coverage. Figure 4.12 upper left shows HCHO at $1x10^{16}cm^{-2}$ precision, approximating the retrieval from a single overpass. The enhancements from burning and urban outflow appear separately distinguishable, however the enhancement between them is not. The lower left of figure 4.12 presents an emulation of the TROPOMI PBL O_3 product, which shows greater O_3 in the northern portion of the region but does not capture the full level of detail. The NO_2 plume from the urban outflow is discernible, though the intensity of the hotspot is underestimated by 50%. Given that the $\sim$hour(s) timescale for the biomass burning event is much less than the resample time of TROPOMI (once per day), an event on the scale of the northern plume may or may not be captured dependending on the timing. Coincidentally, the afternoon overpass time of TROPOMI is close to when the plume was encountered.

4.4.3 TEMPO

The Tropospheric Emissions: Monitoring of POllution (TEMPO) instrument is a NASA instrument to be a hosted payload flown on a commercial geostationary satellite in 2018/2019.

TEMPO will make hourly measurements with 2.1 km/pixel resolution in the north-south direction, 4.7 km/pixel resolution in the east-west direction at the center of the field of regard. O_2, O_3, aerosol, and cloud products will be sampled hourly, including 0-2 km O_3 and free troposphere O_3 for selected target areas. HCHO, CHOCHO, and SO_2 will be reported 3 times/day (hourly samples averaged to attain S/N). Figure 4.12 upper right shows the emulated HCHO retrieval with $4.3x10^{15}cm^{-2}$ precision, representing one of the thrice-daily measurements. TEMPO captures both the biomass burning hotspots and the field of enhancement relative to the background. Figure 4.12 lower right emulates the TEMPO 0-2 km O_3 product and captures essentially all of the major features observed in-situ, including the enhancements and depletion relative to the background. The NO_2 retrieval nicely registers the shape of the plume, and only underestimates the hotspot by 30%. Given the timescale and magnitude of the burning plume, it may feasibly register on one of the scans, and would be dissipated by the next retrieval. In this situation, TEMPO provides data that highlights a temporally-brief pollution event with heavy particulate load near residential regions.

4.5 Conclusions

High-resolution in-situ flight data have been filtered and binned to emulate satellite measurements and estimate the level of detail that may be obtained from the PBL contributions to trace gases' overall satellite column. Based on this county-scale data featuring multiple chemical regimes, it is anticipated that satellite pixels on the scale of 5 x 5 km will resolve many of the chemically-significant features of the spatial distribution of HCHO, O_3, and NO_2 (see §4.3.2). The degree of measurement precision necessary to capture the chemical gradients is identified for each species to provide context for interpreting spectrometer uncertainties (see §4.3.3 and table 4.2). These technical requirements will be achieved with the next generation of satellite instruments, drastically expanding the level of detail (both

spatially and temporally) with which pollutants are mapped (see §4.4). This increases by an order of magnitude the information content from satellite measurements. It is important to anticipate this upcoming influx of data and plan how to utilize it for its full explanative and predictive capacity.

4.6 Future Work

The level of detail expected from TEMPO (see figure 4.12) suggests viability of including feature classification in the satellite data assimilation process. On-the-fly plume boundary identification could be coupled with learning algorithms to provide real-time source localization and emission classification.[50] With the hourly influx of data from TEMPO, techniques from plume modeling and dynamical tracking could connect pollution events over consecutive retrievals.[39][51][10]

4.7 Acknowledgements

The author is grateful for insights from Jessica Gilman and Andy Neuman regarding plume attribution. Thanks to the NOAA Aircraft Operations Center team for providing support for the SONGNEX mission.

5 FUTURE WORK

5.1 Support vector machines

Support vector machines (SVM) are a powerful tool for identifying key data points in enormous data sets, such as those generated by atmospheric measurements.[35][34][43] They are discriminative classifiers that identify decision surfaces in sparse high-dimensional data by finding hyperplanes that maximize the margins between different sets. Application of kernel methods permits generation of nonlinear decisions surfaces, by a variety of mathematical tricks. SVM identifies a small number of points that contain the most information (the“support vectors") and define the boundaries of the margins. Whereas linear regression and neural nets use all data points, SVM focuses on those that are the most difficult to classify. Moving a support vector moves a decision boundary, whereas moving non-support vectors has no impact on classification. This is valuable from an information-theoretic standpoint by highlighting the key data points whose uncertainty most impact the the machine outputs. By transposing the input matrix, we study as well the chemical species and classes of measurements that are most valuable. This can guide experiment design and hone the most important measurements.

5.2 Super Resolution

Super-resolution methods offer a way to better utilize global satellite data of atmospheric chemicals and pollutants such as ozone, particulate matter, sulfur dioxide, and nitrogen dioxide. New satellites are launched every few years, offering more precise and higher-resolution measurements than their predecessors. However, there are periods where multiple generations of instruments are simultaneously in operation. When the satellites' fields of regard overlap, this provides a training data set for improving measurement estimates from the lower-resolution instrument. When the fields of regard do not intersect,

then the older instrument's ongoing measurements can be enhanced by super-resolution. Additionally, these methods can effectively increase the resolution of older measurements taken prior to the newer satellite's launch. Downsampling techniques have been tested in the visual spectrum [3][65] however we will expand this to apply our super-resolution techniques to trace gas retrievals as well. Data at lower processing levels may be valuable for tuning and calibration.

5.3 Markov chain methods & fuzzy comprehensive decision model

Markvo models are capable of predicting air quality based on the empirical topology of regional atmospheric chemistry data.[64] A Markov chain model trained on historical local data predicts air pollution levels, evaluated by a fuzzy comprehensive decision maker. Additionally, local-to-global methods approximate of air quality over a region containing nonuniformly-distributed data points.[63] Together these techniques forecast air quality over an entire region, based on the the current state of the atmosphere (and historical training data) from heterogeneously-distributed geographically-disjoint measurement locations. The performance of the inverse-distance weighted interpolation may be tuned by optimizing scaling coefficients and the distance power, minimizing error as assessed by jackknife resampling. Air quality data are typically accompanied by measurements of the wind vector, which can be taken into account to improve interpolation accuracy. Correlation-weighted interpolation can be included as a corrective term calculated at each station and interpolated over the region. Meteorological factors such as temperature are taken into account by defining fuzzy sets for cold, warm, and hot conditions, then calculating the one-step transition probability matrices for each temperature class.

5.4 Note on cloud computation

Models for analysis of atmospheric big data require great amounts of processing power, and many tasks are well-suited to be broken down and distributed. When possible, models should be written in non-licensed languages (e.g. python rather than MATLAB) to minimize licensing issues with rollout to cloud computation (although there are ways around this using Docker images or resources such as the the Open Science Grid's distributed environment modules). One must consider the specific nature of the problem to be distributed. If each step is dependent on one prior, or one in parallel, one must use high-performance computing (HPC) resources. However a great many problems can be broken down such that individual instances are not dependent on the results of others! In these cases without cross-couplings, high-throughput computing (HTC) is sufficient. At the time of writing, billions of CPU hours per year of HTC resources are available at no cost to academics in the U.S. through networks such as the Open Science Grid. One should immediately consider HTC potential when evaluating large models that incorporate any kind of Monte Carlo sampling, optimization problems, or sweeps of any parameter space. Often, local code can be easily ported to an HTC environment with only a few modifications and the addition of a wrapper script for orientation on a given node. For models whose work can be divided to some degree but include steps that are dependent on previous calculations, HTC implementation can be attained by the use of Directed Acyclic Graphs (DAGs).

BIBLIOGRAPHY

[1] National ambient air quality standards (40 cfr part 50). 1990.

[2] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[3] P. M. Atkinson, E. Pardo-Igúzquiza, and M. Chica-Olmo. *IEEE Geosci. Remote Sens. Lett.*, 46(2):573–80, 2008.

[4] H. Bovensmann, J.P. Burrows, M. Buchwitz, J. Frerick, S. Noël, V.V. Rozanov, K.V., and A.P.H. Goede. *Journal of the Atmospheric Sciences*, 56(2):127–50, 1999.

[5] C.A. Brock, J. Cozic, R. Bahreini, K.D. Froyd, A.M. Middlebrook, A. McComiskey, J. Brioude, O.R. Cooper, A. Stohl, K.C. Aikin, J.A. de Gouw, D.W. Fahey, R.A. Ferrare, R.-S. Gao, W. Gore, J.S. Holloway, G. Hübler, A. Jefferson, D.A. Lack, S. Lance, R.H. Moore, D.M. Murphy, A. Nenes, P.C. Novelli, J.B. Nowak, J.A. Ogren, J. Peischl, R.B. Pierce, P. Pilewskie, P.K. Quinn, T. B. Ryerson, K. S. Schmidt, J. P. Schwarz, H. Sodemann, J. R. Spackman, H. Stark, D. S. Thomson, T. Thornberry, P. Veres, L.A. Watts, C. Warneke, and A.G. Wollny. *Atmos. Chem. Phys.*, 11:2423–53, 2011.

[6] C.A. Brock, F. Schröder, B. Kärcher, A. Petzold, R. Busen, and M. Fiebig. *J. Geophys. Res.*, 105:26555–67, 2000.

[7] J.P. Burrows, M. Weber, M. Buchwitz, V. Rozanov, A. Ladstätter-Weissenmayer, A. Richter, R. DeBeek, R. Hoogen, K. Bramstedt, K.-U. Eichmann, M. Eisinger, and D. Perner. *J. Atmos. Sci.*, 56:151–75, 1999.

[8] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[9] M. Cazorla, G.M. Wolfe, S.A. Bailey, A.K. Swanson, H.L. Arkinson, and T.F. Hanisco. *Atmos. Meas. Tech.*, 8:541–552. doi:10.5194/amt–8–541–2015, 2015.

[10] H. Chen. *Advances in Greedy Algorithms*, ISBN 978-953-7619-27-5, 2008.

[11] W.S. Cleveland, T.E. Graedel, and B. Kleiner. *Atmospheric Environment*, 11(4):357–360, 1967.

[12] E.R. Crosson. *Appl. Phys. B*, 92:403–408, doi:10.1007/s00340–008–3135–y, 2008.

[13] J.P. DiGangi, S.B. Henry, A. Kammrath, E.S. Boyle, L. Kaser, R. Schnitzhofer, M. Graus, A. Turnipseed, J-H. Park, R.J. Weber, R.S. Hornbrook, C.A. Cantrell, R.L. Maudlin III, S. Kim, Y. Nakashima, G.M. Wolfe, Y. Kajii, E.C. Apel, A.H. Goldstein, A. Guenther, T. Karl, A. Hansel, and F.N. Keutsch. *Atmos. Chem. Phys.*, 12:9529–43, 2012.

[14] J. Dixon. *IEEE Transactions on Systems, man, and cybernetics*, 9(10):617–21, 1979.

[15] G. Dominici, R.D. Peng, M.L. Bell, L. Pham, A. McDermott, S.L. Zeger, and J.M. Samet. *JAMA*, 295(10):1127–34, 2006.

[16] N.R. Draper and H. Smith. *Applied Regression Analysis*, pages 307–12, 1998.

[17] S.E. Eftim, J.M. Samet, H. Janes, A. McDermott, and F. Dominici. *Epidemiology*, 19(2):209–16, 2008.

[18] K.-U. Eichmann, K.-P. Heue, and P. Valks. *TROPOMI/S5P Tropospheric Ozone Products ATBD*, 2016.

[19] M. Ester, H.-P. Kriegel, and E. Simoudis J. Han U.M. Fayyad J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, 1996.

[20] M.B. Folette-Cook, K.E. Pickering, J.H. Crawford, B.N. Duncan, C.P. Loughner, G.S. Diskin, A. Fried, and A.J. Jeinheimer. *Atmos. Environ.*, 118:28–44, 2015.

[21] H. Fuchs, W.P. Dubé, B.M. Lerner, N.L. Wagner, E.J. Williams, and S.S. Brown. *Environmental Science & Technology*, VU:doi:10.1021/es902067h, 2009.

[22] S. Funk. Netflix update: Try this at home http://sifter.org/simon/journal/20061211.html. 2006.

[23] P.E. Gill and W. Murray. *SIAM Journal on Numerical Analysis*, 15(5):977–92, 1978.

[24] S.M.K. Heris. Implementation of dbscan clustering in matlab. 2015.

[25] R. Holzinger, C. Warneke, A. Hansel, A. Jordan, W. Lindinger, D.H. Scharffe, G. Schade, and P.J. Crutzen. *Geophysical Research Letters*, 26(8):1161–1164, 1999.

[26] P.K. Hopke, C. Liu, and D.B. Rubin. *Journal of the International Biometric Society*, 57(1):22–33, 2001.

[27] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen. *Atmos. Environ*, 38:2895–907, 2005.

[28] J. Kaiser, G.M. Wolfe, B. Bohn, S. Broch, H. Fuchs, L.N. Ganzeveld, S. Gomm, R. Häseler, A. Hofzumahaus, F. Holland, J. Jäger, X. Li, I. Lohse, K. Lu, A.S.H. Prévôt, F. Rohrer, R. Wegener, R. Wolf, T. F. Mentel, A. Kiendler-Scharr, A. Wahner, and F. N. Keutsch. *Atmos. Chem. Phys.*, 15:1289–1298, 2015.

[29] L. Kaufman and P.J. Rouseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., 1990.

[30] A.R. Koss, C. Warneke, B. Yuan, M.M. Coggon, P.R. Veres, and J.A. de Gouw. *Atmospheric Measurement Techniques*, 9:2909–25, 2016.

[31] M. Kurucz, A.A. Benczúr, and K. Csalogány. *KDDCup.07*, pages ACM 978–1–59593–834–3/07/0008, 2007.

[32] W.A. Lahoz, V.-H. Peuch, J. Orphal, J.-L. Attié, K. Chance, X. Liu, D. Edwards, H. Elbern, J.-M. Flaud, M. Claeyman, and L. El Amraoui. *Bull. Amer. Meteor. Soc.*, 93:221–33, 2012.

[33] S.P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[34] W.-Z. Lu and D. Wang. *Science of the Total Environment*, 395(2-3):109–16, 2008.

[35] W.-Z. Lu and W.-J. Wang. *Chemosphere*, 59(5):693–701, 2005.

[36] S.T. Martin, P. Artaxo, L.A.T. Machado, A.O. Manzi, R.A.F. Souza, C. Schumacher, J. Wang, M.O. Andreae, H.M.J. Barbosa, J. Fan, G. Fisch, A.H. Goldstein, A. Guenther, J.L. Jimenez, U. Pöschl, M.A. Silva Dias, J.N. Smith, and M. Wendisch. *Atmos. Chem. Phys.*, 16:4785–97, 2016.

[37] D. L. Mauzerall and X. P Wang. *Ann. Rev. Energ. Environ*, 26:237–268, 2001.

[38] G. McLachlan and D. Peel. *Finite Mixture Models*. 2000.

[39] L. Merino and A. Ollero. *28th Annuan Conference of the IEEE*, 7754740:10.1109/IECON.2002.1185316, 2002.

[40] J.J. Moré and D.C. Sorensen. *SIAM J. Sci. Stat. Comput.*, 4:553–572, 1983.

[41] S. Moshenberg and U. Lerner. *Environ. Syst. Res.*, 4:26, 2015.

[42] NOAA. https://www.esrl.noaa.gov/csd/projects/songnex/. *Shale Oil and Natural Gas Nexus homepage*.

[43] S. Osowski and K. Garanty. *Engineering Applications of Artificial Intelligence*, 20(6):745–55, 2007.

[44] H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 36:3336–3341, 2009.

[45] J. Peischl, T.B. Ryerson, J.S. Holloway, M. Trainer, A.E. Andrews, E.L. Atlas, D.R. Blake, B.C. Daube, E.J. Dlugokencky, M.L. Fischer, A.H. Goldstein, A. Guha, T. Karl, J. Kofler, E. Kosciuch, P.K. Misztal, A.E. Perring, I.B. Pollack, G.W. Santoni, J.P. Schwarz, J.R. Spackman, S.C. Wofsy, and D.D. Parrish. *J. Geophys. Res.*, 117:D00V25, doi:10.1029/2012JD017994, 2012.

[46] I.B. Pollack, B.M. Lerner, and T.B. Ryerson. *Journal of Atmospheric Chemistry*, 65:2-3:111–25, 2011.

[47] P.J. Rouseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[48] D.B. Rubin. *Biometrika*, 63(3):581–92, 1976.

[49] T.B. Ryerson, L.G. Huey, K. Knapp, J.A. Neuman, D.D. Parrish, D.T. Sueper, and F.C. Fehsenfeld. *Journal of Geophysical Research-Atmospheres*, 140:D5:5483–92., 1999.

[50] S. Sahyoun. Plume source localization and boundary prediction.

[51] S.S. Sahyoun, S.M. Djouadi, and H. Qi. *2010 American Control Conference*, 978-1-4244-7425-7/10:2915–2920, 2010.

[52] J.M. Samet, F. Dominici, F.C. Curriero, I. Coursac, and S.L. Zeger. *N Engl J Med*, 343:1742–1749, 2000.

[53] D.J. Seidel, C.O. Ao, and K. Li. *Journal of Geophysical Research*, 115:D16113, 2010.

[54] R. Spurr, D. Loyola, and C. Lerot. *S5P/TROPOMI Total Ozone ATBD*, 2016.

[55] D.M. Stieb, R. C. Beveridge, J. R. Brook, M.A.R.C. Smith-Doiron, R.T. Burnett, R.E. Dales, S. Beaulieu, S. Judek, and A. Mamedov. *J. Exp. Anal. Environ. Epidemiol.*, 10:461–477, 2000.

[56] G. Sugiyama and J.S. Nasstrom. Methods for determining the height of the atmospheric boundary layer. *LLNL report*.

[57] J.H.G.M. van Geffen, K.F. Boersma, H.J. Eskes, J.D. Maasakkers, and J.P. Veefkind. *TROPOMI ATBD of the total and tropospheric NO_2data products*, 2016.

[58] K. Vijayaraghavan, H.E. Snell, and C. Seigneur. *Environ. Sci. Technol.*, 42(22):8187–92, 2008.

[59] C. Willmott. *Bulletin American Meteorological Society*, 63(11):1309–13, 1982.

[60] C. Willmott, S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe. *J. Geophys. Res.*, 90(C5):8995–9005, 1986.

[61] G.M. Wolfe, J. Kaiser, T.F. Hanisco, F. N. Keutsch, J. A. de Gouw, J.B. Gilman, M. Graus, C.D. Hatch, J. Holloway, L.W. Horowitz, B.H. Lee, B.M. Lerner, F. Lopez-Hilifiker, , J. Mao, M.R. Marvin, J. Peischl, I.B. Pollack, J.M. Roberts, T.B. Ryerson, J.A. Thornton, P.R. Veres, and C. Warneke. *Atmos. Chem. Phys.*, 16:2597–2610, 2016.

[62] B. Yuan, A. Koss, C. Warneke, J.B. Gilman, B.M. Lerner, H. Stark, and J.A. de Gouw. *Atmospheric Measurement Techniques*, 9:2735–52, 2016.

[63] C. Zhu, M.P. Krawiec-Thayer, and A.H. Assadi. Stochastic methods for air quality prediction. *4th Annual Conf. on Comp. Sci. & Comp. Intelligence*, 2017.

[64] C. Zhu, M.P. Krawiec-Thayer, G. Huijing, and A.H. Assadi. Topological methods for modeling stochastic events in atmospheric chemistry. *American Mathematical Society (AMS) Fall Southeastern Sectional Meeting*, 2017.

[65] S. Peleg Zomet. *IEEE Workshop on Applications of Computer Vision*, WACV'02:0–7695–1858–3/02, 2002.

[66] P. Zoogman, X. Liu, R.M. Suleiman, W.F. Pennington, D.E. Flittner, J.A. Al-Saadi, B.B. Hilton, D.K. Nicks, M.J. Newchurch, J.L. Carr, S.J. Janz, M.R. Andraschko, A. Arola,

B.D. Baker, B.P. Canova, C. Chan Miller, R.C. Cohen, J.E. Davis, M.E. Dussault, D.P. Edwards, J. Fishman, A. Ghulam, G. González Abad, M. Grutter, J.R. Herman, J. Houck, D.J. Jacob, J. Joiner, B.J. Kerridge, J. Kim, N.A. Krotkov, L. Lamsal, C. Li A. Lindfor, R.V. Martina, C.T. McElroy, C. McLinden, V. Natraj, D.O. Neil, C.R. Nowlan, E.J. O'Sullivan, P.I. Palmer, R.B. Pierce, M.R. Pippin, A. Saiz-Lopez, R.J.D. Spurr, J.J. Szykman, O. Torres, J.P. Veefkind, B. Veihelmann, H. Wang, J. Wang, and K. Chance. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 186:17–39, 2016. doi:10.1016/j.jqsrt.2016.05.008.