STATISTICAL METHODS FOR THE INTERPRETATION, PREDICTION, AND LOCALIZATION OF REMOTELY SENSED ATMOSPHERIC POLLUTANTS

by

William S. Daniels

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Applied Mathematics and Statistics).

Golden, Colorado

Date _____

Signed: _____

William S. Daniels

Signed: _____

Dr. Dorit Hammerling
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____

Dr. Greg Fasshauer
Professor and Department Head
Department of Applied Mathematics and Statistics

ABSTRACT

We present two statistical modeling efforts that seek to address pressing environmental issues through the use of remotely sensed data. The first study is motivated by the extreme fire seasons now commonly experienced across the globe (e.g. the 2015 Indonesian forest fires and the 2019/2020 Australian bush fires). We develop interpretable models for remotely sensed carbon monoxide, a proxy for fire intensity in the Southern Hemisphere, fit using a flexible regularization framework. These models are parsimonious by design, allowing for scientific insight into the primary climate drivers of fire season intensity in different regions. The models have good predictive skill at considerable lead times, making them a useful tool for predicting upcoming fire season intensity. The second study is motivated by a growing dependence on natural gas for energy in the United States. Methane (the primary component of natural gas) burns cleaner than coal and oil but is a potent greenhouse gas. Therefore, limiting emissions during natural gas production is essential if it is to be considered a cleaner alternative to other fossil fuels. With the goal of localizing small-scale emissions, we develop a hierarchical spatial model for estimating methane concentrations on a fine grid given coarsely pixelated satellite observations. We apply our model to a satellite overpass of the Denver-Julesburg (DJ) Basin (located in northeast Colorado) to demonstrate its effectiveness. We use conditional simulation for uncertainty quantification and inferences related to emissions monitoring.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAO . . . . . Antarctic Oscillation

BIC . . . . . Bayesian Information Criterion

CO . . . . . Carbon Monoxide

DMI . . . . . Dipole Mode Index

EBIC . . . . . Extended Bayesian Information Criterion

ENSO . . . . . El Nino-Southern Oscillation

IOD . . . . . Indian Ocean Dipole

LASSO . . . . . Least Absolute Shrinkage and Selection Operator

MJO . . . . . Madden-Julian Oscillation

MLE . . . . . Maximum Likelihood Esimate

MOPITT . . . . . Measurements Of Pollution In The Troposphere

MSEA . . . . . Maritime Southeast Asia

OLR . . . . . Outgoing Longwave Radiation

OMI . . . . . Ozone Monitoring Instrument

RAMP . . . . . Regularization Algorithm under Marginality Principle

SAM . . . . . Southern Annular Mode

TROPOMI . . . . . Tropospheric Monitoring Instrument

TSA . . . . . Tropical South Atlantic

ACKNOWLEDGMENTS

Thanks to all that have supported me during my time at Mines. To Dorit, for the continued support and excellent guidance, for deciding to work with a physics undergrad with zero stats background, and for always finding my typos. To Doug, for being the first person to truly show me the connections between theory and data. To Rebecca, for introducing me to the fascinating field of atmospheric chemistry and always answering my questions. To Morgan, for giving me the opportunity to work on projects that can make a real difference.

To my parents and Kate for their constant love and support during stressful times. And finally, to Hugh and Michelle Harvey for giving me the flexibility to switch fields and work on what I'm passionate about and for showing me what it means to pay it forward.

CHAPTER 1

INTRODUCTION

Satellite remote sensing has produced a wealth of data with wide ranging environmental applications, and we present two specific examples in this thesis. Before discussing these studies, however, we provide a brief overview of satellite remote sensing, the types of data it produces, and some challenges associated with using these data. We believe that this will provide useful context for the two specific applications presented in Chapters 2 and 3.

Satellite remote sensing (or just "remote sensing" for the remainder of this text) refers to the use of satellite-based instruments to gather information about an object or phenomenon. Specifically, we focus on remote sensing of Earth. These remote sensing instruments collect data by measuring the intensity of electromagnetic radiation reflected by the Earth's surface. Passive sensors detect the reflection of naturally produced light (e.g. sunlight), while active sensors detect the reflection of an artificially produced light (e.g. a laser). Hyperspectral imaging is one type of remote sensing that is often used to measure the concentration of trace gasses in the atmosphere. As the reflected light interacts with these gasses, such as carbon monoxide or methane, the spectra of the light are altered before being recorded by the satellite-based spectrometers. Algorithms are then able to analyze these spectra and determine the concentrations of the gasses based on their known absorption properties. The result of this process is the "retrieval." We direct the interested reader to Levelt *et al.* (2006), Veefkind *et al.* (2012), and Hu *et al.* (2016) as a starting point for more information.

Remotely sensed data has a number of features related to the satellite platform from which it is measured. The most obvious is simply the location of the measurements. Unlike ground-based sensors that are located at fixed points in space, such as the 23 instruments in the Total Carbon Column Observing Network (TCCON) (Wunch *et al.*, 2011), satellite-based sensors continuously orbit the planet. The result is that each retrieval observes a slightly different portion of the Earth's surface. An obvious benefit to this type of measurement is that a single satellite-based instrument can monitor the entire globe. However, it also complicates the analysis of these data. First, it introduces two extra dimensions in addition to time (latitude and longitude). Second, it makes it unlikely for two observations to occur at exactly the same point in space. As a result, it is often necessary to spatially aggregate observations that occur close together when studying a specific region.

A related feature of remotely sensed data is the resolution of the observations. Satellite instruments gather data by repeatedly imaging stretches of the Earth's surface for short durations of time. This results

in a sequence of rectangular observations, with the across-track length defined by the viewing angle of the instrument and the along-track length defined by the time used to create the image. The across-track dimension is broken up according to the number of pixels in the imaging system. The result of this measurement strategy is a number of observation "footprints," where the footprints are the projection of the imaging pixels onto the surface of the Earth. The trace-gas observation associated with each footprint is a function of the gas within the column extending from the boundary of that footprint. The footprints exhibit different shapes depending on their relative position to the nadir (straight down location) of the satellite. Specifically, as the rectangular observations are projected onto the curved surface of the Earth, the footprints towards the ends of the rectangle get stretched out, while the pixels closer to nadir are less distorted. See Veefkind *et al.* (2012) for further details.

While these footprints can be approximated by their center point, the observation is a function of the target gas within the volume of the column extending from the footprint. These footprints can be largely ignored if aggregating data over large spatial regions but are important to consider when analyzing small regions closer in scale to the size of the footprints.

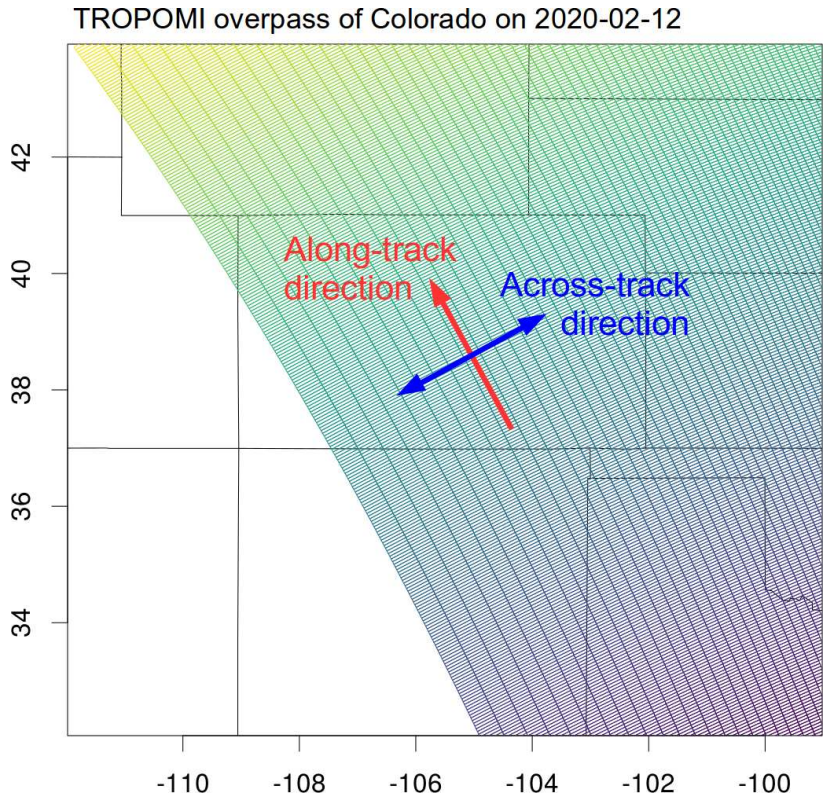

Figure 1.1 Example overpass of the TROPOMI satellite instrument to demonstrate the pixelated nature of the observations. Along-track and across-track directions are shown. Each box within the grid shows a footprint corresponding to an observation. Color represents the time of measurement, with cooler colors occurring before the hotter colors.

The satellite overpass of Colorado shown in Figure 1.1 is an example of this pixelated measurement property. These observations are from the TROPOMI instrument onboard the Copernicus Sentinel-5 Precursor satellite (discussed further in Chapter 3). Each box within the colored grid shows a footprint corresponding to an observation. Here color represents the time of measurement, with cooler colors occurring before the hotter colors. The along-track direction is shown with a red arrow, and the across-track direction is shown with a blue arrow. Note that these footprints are very elongated because they are located at the end of the two dimensional strip described above.

The final feature that we will discuss is the time of measurement. Satellites can be placed in different orbits around the Earth, with each resulting in a different path traced along the Earth's surface as the satellite orbits. Satellites are often placed in a sun-synchronous orbit when global coverage is desired, which results in measurements occurring at the same local mean solar time at every point across the globe. In other words, the satellites typically observe each point on the globe at around the same local time each day. This is important for consistency, as atmospheric gasses often have diurnal (i.e. daily) cycles that affect their concentrations.

Having broadly discussed these properties of satellite remote sensing, we now briefly introduce the two studies presented in Chapters 2 and 3. Both studies involve statistical models that seek to address pressing environmental issues through the use of satellite data. The goals of each study can be summarized as follows:

1. Explore the relationship between climate variability and fire season intensity. To this end, we develop interpretable models for remotely sensed carbon monoxide (a proxy for fire intensity in the Southern Hemisphere).

2. Explore the capabilities of remotely sensed methane as a tool for localizing small scale emissions (both fugitive and operational) from oil and gas production. For this application, we develop a spatial model that predicts methane concentrations on a fine grid given the coarsely pixelated satellite observations.

For the first study, we develop a lagged, multiple linear regression model that is fit with a flexible regularization framework. We average carbon monoxide observations onto a regional scale, meaning that we can essentially ignore the footprint of each observation. For the second study, we develop a spatial model fit via maximum likelihood. We make predictions at a resolution finer than the satellite footprints, and therefore we take into consideration the shape of the footprints when fitting the model. We leave further introduction to these projects for their respective chapters, where we discuss in detail the current literature, our research goals, and the methods we have developed.

The rest of this document is laid out as follows. In Chapters 2 and 3, we discuss the two studies briefly introduced above. Within each of these chapters, we discuss the current literature, our research goals, the remotely sensed data utilized, the methods we develop, and our results. In Chapter 4, we briefly summarize our work and make connections between the two projects.

CHAPTER 2
INTERPRETABLE MODELS FOR THE ANALYSIS AND PREDICTION OF FIRE SEASON INTENSITY

2.1 Introduction

There have been many devastating fire seasons in recent years across the globe. Two examples are the 2015 Indonesian forest fires and the 2019/2020 Australian bushfires. In fact, the 2015 fire season in Indonesia was the most severe fire activity in the region since the NASA Earth Observing System satellites came online in the early 2000s. During this two month event, thick smoke blanketed Sumatra and Kalimantan and millions of people were exposed to hazardous air quality (Field *et al.*, 2016). This motivates the need for advanced predictions of fire season intensity, as they would give countries like Indonesia time to better prepare for these extreme fire events.

The relationship between fire and climate has been extensively studied. Fire intensity and burned area are related to the amount, type, and dryness of available fuel, all of which respond closely to water conditions driven by climate variability (van der Werf *et al.*, 2008). This relationship is complex and varies across the different regions of the globe. For instance, drought conditions were found to increase fire potential in southern Africa, but decrease fire potential in northern Africa (Andela & Van Der Werf, 2014).

Climate modes, such as the El Nino Southern Oscillation (ENSO), capture variability in the global climate system. Studies have used these climate modes to help explain the complex relationship between climate and fire, often using a regression modeling framework. ENSO has been found to influence fires in North America (Shabbar *et al.*, 2011), Maritime Southeast Asia (Fuller & Murphy, 2006; Reid *et al.*, 2012), the Amazon (Alencar *et al.*, 2011), and Africa (Andela & Van Der Werf, 2014). However, studies have found that fire behavior is often related to several distinct climate modes (Andreoli & Kayano, 2006; Chen *et al.*, 2016; Saji & Yamagata, 2003). In fact, Cleverly *et al.* (2016) finds that the interactions between the ENSO, IOD, and AAO climate modes are particularly important for explaining drought and rainfall in Australia, which in turn are major drivers of fire activity. This indicates that fire behavior is affected not only by the isolated influence of multiple modes, but also by their interactions (i.e. whether or not the modes are in phase).

In addition to identifying the influential climate modes for fire behavior in a given region, studies such as Chen *et al.* (2016) and Wooster *et al.* (2012) identify lead times that correspond to the maximum predictive skill of the climate modes being studied. Similarly, Shawki *et al.* (2017) examines how far in

advance the 2015 fire event in Indonesia can be predicted using climate based models, finding that lead times of up to 25 weeks can still provide useful predictions.

These fire-climate connections have been previously studied using satellite observations of fire properties (e.g. Ceccato *et al.* (2010), Wooster *et al.* (2012), and Chen *et al.* (2016)). The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments onboard the Terra and Aqua satellites provide fire count data for each overpass as well as a burned area data product (Giglio *et al.*, 2016, 2018). However, using fire counts or burned area directly presents a number of challenges. Fire counts ignore differences in fire size and intensity, and burned area products potentially miss small fires, underground peat fires, and fires obscured by smoke (although significant improvements in this regard have been made with the most recent product) (Giglio *et al.*, 2018; Shawki *et al.*, 2017).

One alternative is to model atmospheric carbon monoxide (CO) instead of fire counts or burned area directly. CO is produced by incomplete combustion from biomass burning, fossil fuel use, and indirectly by photochemistry (Buchholz *et al.*, 2018; Holloway *et al.*, 2000), and its link to fires is well established (Edwards *et al.*, 2006a). In fact, biomass burning is the primary source of atmospheric CO variability in the Southern Hemisphere. Thus CO anomalies are a useful proxy for fire intensity (Voulgarakis *et al.*, 2015). Compared to the study of fire counts and burned area, less research has gone into the connection between atmospheric CO and climate variability. Because CO variability in the Southern Hemisphere is closely linked to biomass burning, it also responds to variability in the climate. Furthermore, modeling CO provides information on atmospheric pollutants concurrently with information on fire intensity.

Edwards *et al.* (2006b) found that CO observations from the Measurement of Pollution in the Troposphere (MOPITT) instrument is correlated with ENSO. Buchholz *et al.* (2018) expanded on Edwards *et al.* (2006b), finding that atmospheric CO anomalies in a number of Southern Hemisphere regions are correlated to four different climate modes (including ENSO) and that the interactions between these climate modes is important in explaining atmospheric CO anomalies. Buchholz *et al.* (2018) used month-averaged CO and climate mode data and separated lag and variable selection into two computational steps.

We also focus on the connection between atmospheric CO and climate variability, expanding on Buchholz *et al.* (2018) via the following advancements.

- We create models using week-averaged data, rather than month-averaged data, significantly increasing predictive skill.
- We include a proxy for a fifth climate mode, the Madden-Julian Oscillation (MJO), resulting in models that are better able to capture extreme CO anomalies in Maritime Southeast Asia.

- We develop a more flexible model fitting framework that performs variable and lag selection simultaneously using regularization. This allows for multiple lags of a single climate mode in the statistical models.
- We develop a framework for assessing the stability of selected model terms. This gives weight to the scientific interpretation of the selected model terms and ultimately improves model interpretability.
- We incorporate an option for setting the minimum and maximum lags allowed in the statistical models, making it possible to set the desired lead time of model predictions.

With these advancements in mind, we focus on two main research goals:

1. Create interpretable models that can be used to draw scientific conclusions about the connection between climate and atmospheric chemistry.
2. Create models with a high level of predictive skill that can be used to predict fire season intensity reasonably far in advance.

The rest of this chapter is laid out as follows. In Sections 2.2 and 2.3, we describe the data and the statistical model, respectively. In Section 2.4, we discuss the model fitting framework we have developed for this application. In Sections 2.5 and 2.6 we discuss how we use this modeling framework to address the two research goals listed above. Finally, we summarize this work in Section 2.7.

2.2 Observational Data Sets

2.2.1 Response Variable

For the response, we use carbon monoxide column-averaged volume mixing ratios (referred to as simply CO) from the MOPITT instrument onboard the Terra satellite. The units of column-averaged volume mixing ratios are parts per billion by volume (ppb). Using column-averaged volume mixing ratios instead of total column CO removes dependence on surface topography and pressure changes.

MOPITT has complete Earth coverage about every three days with a footprint size of 22×22 km. We use the latest retrieval algorithm (V8), which has been validated in Deeter *et al.* (2019). To reduce systematic and random error, we select daytime, land-only retrievals from the thermal infrared (TIR) product. Daytime retrievals have a higher sensitivity to CO than nighttime retrievals due to higher thermal contrast during the day. Land-only retrievals have less error than water-only retrievals because MOPITT is more sensitive to CO over land scenes. Finally, the TIR has less random error than the near-infrared or multispectral products. See Buchholz *et al.* (2018), Deeter *et al.* (2007), and Deeter *et al.* (2014) for details.

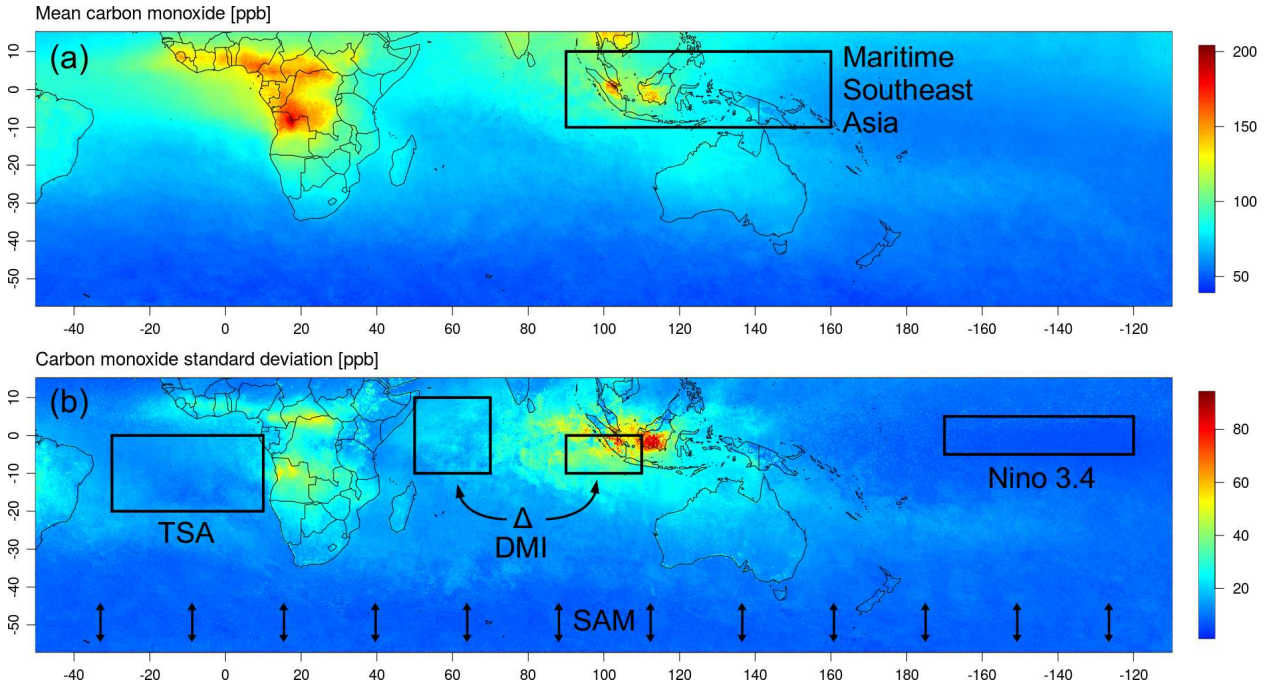

Figure 2.1 (a) Average CO during 2015 as measured by MOPITT with the Maritime Southeast Asia (MSEA) response region overlaid in black. (b) CO standard deviation during the same time period with the spatial range of influence of the four climate mode indices overlaid in black.

We aggregate CO observations into a single biomass burning region in the Southern Hemisphere. We focus on Maritime Southeast Asia (MSEA) in this paper, but it should be noted that this methodology could be applied to other regions as well (such as Southeast Australia - the region that experienced severe bushfires in 2019/2020). We focus on MSEA because it is a biomass burning region that experiences significant CO anomalies (ie. CO concentrations well above average). Figure 2.1(a) shows the MSEA response region overlaid on the average CO during 2015.

We create a weekly time series for the MSEA region by averaging all of the observations falling within the region boundaries (see Figure 2.1(a)) for each week. This time series contains 18 years of data, from 2001 to 2019. We remove the annual seasonal cycle from the weekly time series so that our models are better able to capture the anomalous CO observations corresponding to large burn events. We compute the seasonal cycle by taking an average over the 18 years of data for each week. This climatological average is then subtracted from the weekly time series to create CO anomalies.

Finally, since we are interested in using CO as a proxy for fires, we only model the anomalies during fire season in the Southern Hemisphere, defined here as September through December. This time range was selected based on results from Buchholz *et al.* (2018) showing that these months captured most of the atmospheric CO variability in the MSEA region. The CO anomalies during fire season are used as the

response variable in our models. Figure 2.2 shows the weekly CO observations, climatological average, and resulting anomalies for the MSEA region. Note that data for the entire year is plotted, even though we only model September through December.

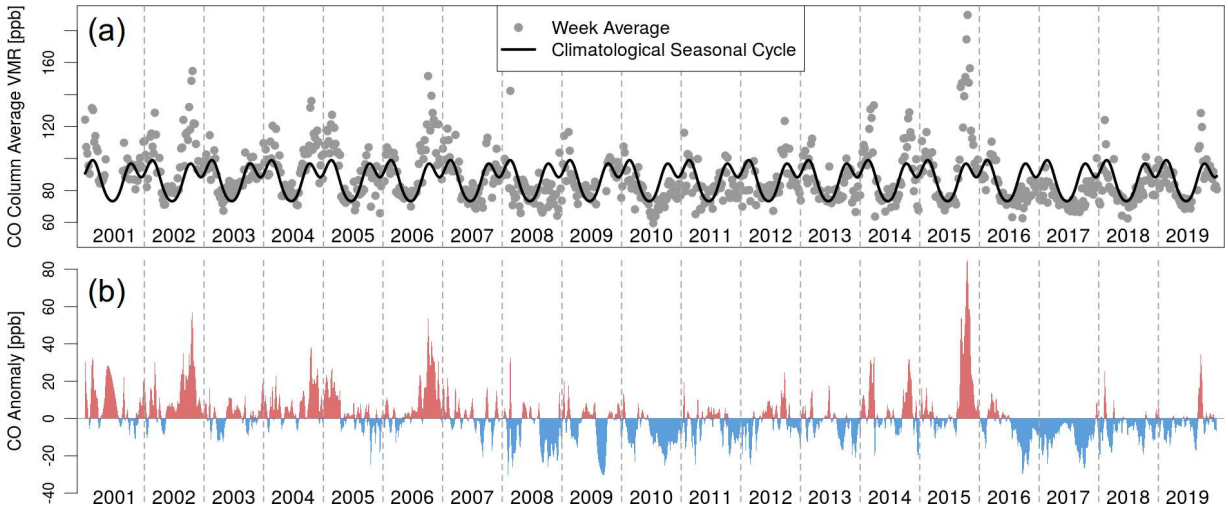

Figure 2.2 (a) Weekly CO observations for the MSEA response region and the climatological average created by averaging each week over the 18 year time series. (b) Anomalies resulting from the difference between the weekly observations and the climatological average.

2.2.2 Predictor Variables

We are interested in connections between atmospheric CO and climate variability. Climate modes are large scale patterns that capture variation in temperature, wind, or other aspects of climate over certain spatial regions. A well known example is the El Nino Southern Oscillation (ENSO), which captures quasi-periodic variability in sea surface temperature and wind in the Pacific Ocean (Neelin *et al.*, 1998; Trenberth, 2013). Climate indices are metrics that quantify the state of climate modes. Multiple climate indices exist for each climate mode. For instance, the Nino 1.2, Nino 3.4, and Southern Oscillation Index (SOI) are all indices that describe the state of ENSO (NOAA OOPC, 2021).

Here we consider four climate modes that represent variability in the major ocean basins of the Southern Hemisphere and tropics. The El Nino Southern Oscillation (ENSO) represents the Pacific Ocean, the Indian Ocean Dipole (IOD) represents the Indian Ocean, the Tropical South Atlantic (TSA) represents the southern Atlantic Ocean, and the Antarctic Oscillation (AAO) represents the Southern Ocean. These climate modes were selected to match the analysis in Buchholz *et al.* (2018).

For predictor variables, we select a single climate mode index to represent each of these climate modes. To represent the ENSO, we use the Nino 3.4 index defined in Bamston *et al.* (1997). To represent the TSA, we use the Tropical South Atlantic Index defined in Enfield *et al.* (1999). These two indices are calculated

using sea surface temperature (SST) anomalies in the regions shown in Figure 2.1(b) labeled as Nino 3.4 and TSA, respectively. To represent the IOD, we use the Dipole Mode Index (DMI) defined in Saji *et al.* (1999). This index is calculated from SST gradients between the two regions shown in Figure 2.1(b) labeled as DMI. To represent the AAO, we use the Southern Annular Model (SAM) index defined in Thompson & Wallace (2000). This index captures Antarctic atmospheric circulation described by the poleward shift of westerly winds. This index is calculated by projecting observational height anomalies at 700 hPa and poleward of -20 degrees latitude onto the leading empirical orthogonal function of the National Centers for Environmental Prediction and National Center for Atmospheric Research reanalysis (Kalnay *et al.*, 1996; Kistler *et al.*, 2001). The spatial extent of this index is shown in Figure 2.1 via the arrows labeled SAM. We expect a relationship between these indices and CO, as each index affects regional climate (e.g. rainfall), which in turn affects drought, fire, and ultimately CO loading.

In addition to these four indices, we also want to include variability captured by the Madden-Julian Oscillation (MJO) climate mode. This climate mode broadly describes the eastward propagation of a convection cell that forms off the east cost of Africa and dissipates in the Pacific Ocean (Madden & Julian, 1972). The MJO is the dominant mode of intraseasonal variability in the tropics (Madden & Julian, 1994) and has been shown to increase or decrease the probability of extreme rain events by over 20% in the MSEA region depending on its phase (Xavier *et al.*, 2014). However, unlike the other climate modes included in this study, the most common MJO index is described by the two primary empirical orthogonal functions (EOFs) resulting from a number of climate variables (Wheeler & Hendon, 2004). Included in the EOF analysis is outgoing longwave radiation (OLR), a metric that describes how much energy is leaving the atmosphere. Low OLR values indicate the presence of clouds, and hence a higher likelihood of rainfall.

To capture the variability described by the MJO in our models, we use OLR anomalies instead of the two primary EOFs from Wheeler & Hendon (2004). This is done to better accommodate a linear regression framework. The phase of the MJO depends on both EOFs simultaneously, which could be included in a linear regression model by using a main term for both EOFs and their interaction. However, this introduces three covariates (and hence three coefficient estimates) to capture a single physical phenomenon. This makes it harder to correctly model the contribution of the MJO and hinders model interpretation. Using OLR anomalies in the MSEA region as a proxy for the MJO provides a single metric that captures the presence of the convection cell described by the MJO. We believe and have confirmed through preliminary testing that this proxy, while losing some of the information contained in the two MJO EOFs, is better suited for a regression analysis. OLR values are aggregated over the same spatial region that defines the MSEA response region shown in Figure 2.1, and anomalies are created in the same manner as the CO anomalies described in the previous section. We consider the inclusion of the OLR as a proxy for

the MJO a major advancement over the models presented in Buchholz *et al.* (2018), and we demonstrate the benefit of including the OLR proxy in Section 2.6.1.

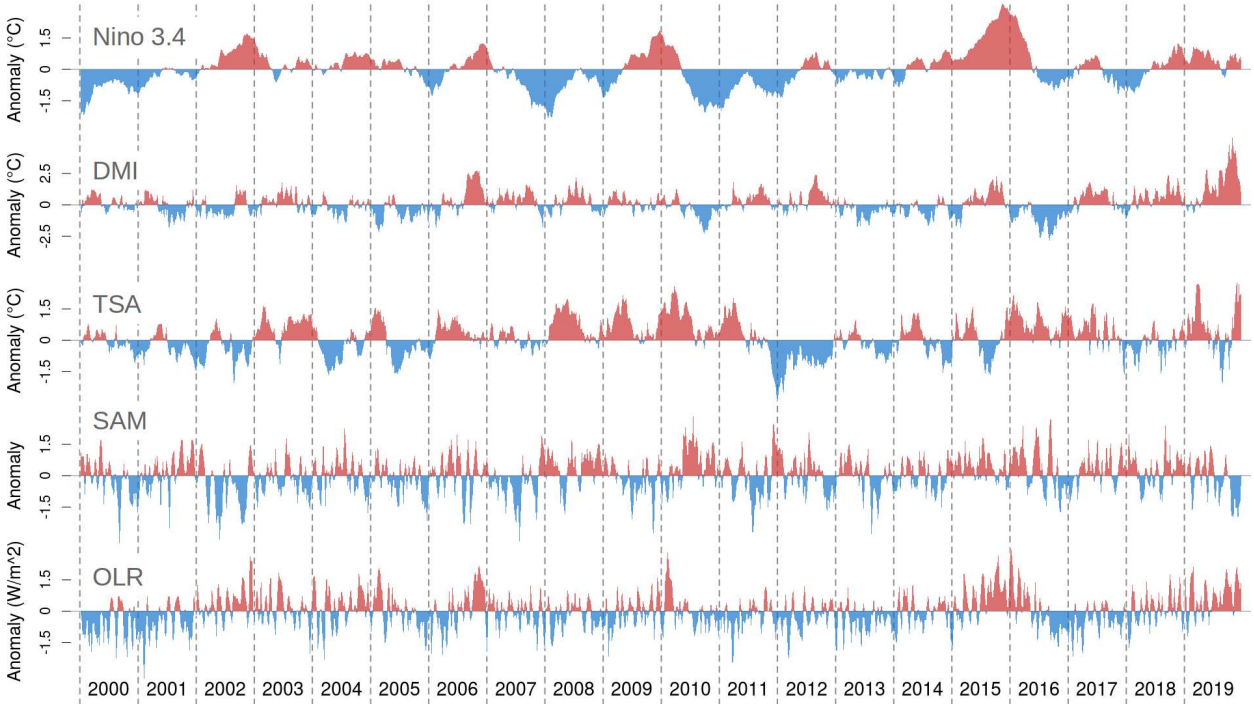

Figure 2.3 Time series of the five climate mode indices used as predictor variables in this study. Note that OLR is used as a proxy index for the MJO.

Figure 2.3 shows the weekly time series for each climate mode index used as a predictor variable in this study. Some of the indices have both high and low frequency components. This is most obvious in the SAM and OLR. We believe that the high frequency component of the OLR captures oscillatory movement of the convection cell described by the MJO. The movement of this convection cell has a period of 30 to 90 days, which is closely aligned with the period of the high frequency component of the OLR.

The climate mode index data used in this study are publicly available. The source of each index (or proxy index in the case of the MJO) is listed in Table 2.1.

Table 2.1 Climate mode indices used in this study with links to their sources. Note that we use OLR as a proxy index for the MJO.

Climate Mode	Associated Index	Source
ENSO	Nino 3.4	NOAA OOPC (2021)
IOD	Dipole Mode Index (DMI)	NOAA OOPC (2021)
TSA	Tropical South Atlantic (TSA)	NOAA OOPC (2021)
AAO	Southern Annular Mode (SAM)	NOAA CPC (2021)
MJO	Outgoing Longwave Radiation (OLR)	NOAA PSL (2021)

2.3 Multiple Linear Regression Model

We use lagged multiple linear regression to model the relationship between CO anomalies and climate mode indices. We include first order interaction terms to capture the interconnected nature of the global climate system. Buchholz *et al.* (2018) found that these interaction terms were highly significant in explaining CO variability. We also include squared terms to capture potential non-linear relationships between the mean CO response and the climate mode indices. These terms are not included in the models in Buchholz *et al.* (2018), and we believe that they allow the models to capture more complex relationships. For a given response region, we assume that

$$CO(t) = \mu + \sum_{k} a_k \cdot \chi_k(t - \tau_k) + \sum_{i,j} b_{ij} \cdot \chi_i(t - \tau_i) \cdot \chi_j(t - \tau_j) + \sum_{l} c_l \cdot \chi_l(t - \tau_l)^2 + \epsilon, \tag{2.1}$$

where $CO(t)$ is the CO anomaly at time t, μ is a constant mean displacement, a_k, b_{ij}, and c_l are coefficients, χ are the climate indices, τ is the lag value for each index in months, ϵ is a random error component, and k,i, and j iterate over the number of climate modes used in the analysis. We consider lags between one and 52 weeks for each index. We do not consider zero week lags so that our models can be used for prediction.

We do not expect the high frequency variability in the climate mode indices to have a large effect on the amount, type, and dryness of available fuel far into the future. This is because short lasting anomalies (either positive or negative), while potentially having an important short term impact, do not last long enough to drastically alter large scale fuel reserves. Therefore, we want covariates with longer lags to capture progressively lower frequency components of the climate indices.

To accomplish this, we apply more smoothing to the climate mode indices as the length of their lag in the statistical model increases. Specifically, we employ the following smoothing strategy. We do not smooth the indices for lags below four weeks, as we want to capture as much high frequency signal as possible from these very short term relationships. For lags between four and 52 weeks, we use a Gaussian kernel to smooth the indices, with the bandwidth value increasing every four weeks. To select bandwidth values, we first found the bandwidth that seemed to best capture the long term trend in the climate indices. This was then set as the maximum bandwidth and a continuous sequence of bandwidth values was created between no smoothing and this maximum value.

Figure 2.4 shows every other level of smoothing applied to the climate indices over two years of data. The black curve is the original weekly climate index time series, which is used for lags one through three. The colored curves show every other level of smoothing up to the maximum smoothing applied to lags of one year. Note that the vertical axis has been omitted from Figure 2.4 for visual clarity since its purpose is

solely to show the relative levels of smoothing applied to each climate index.

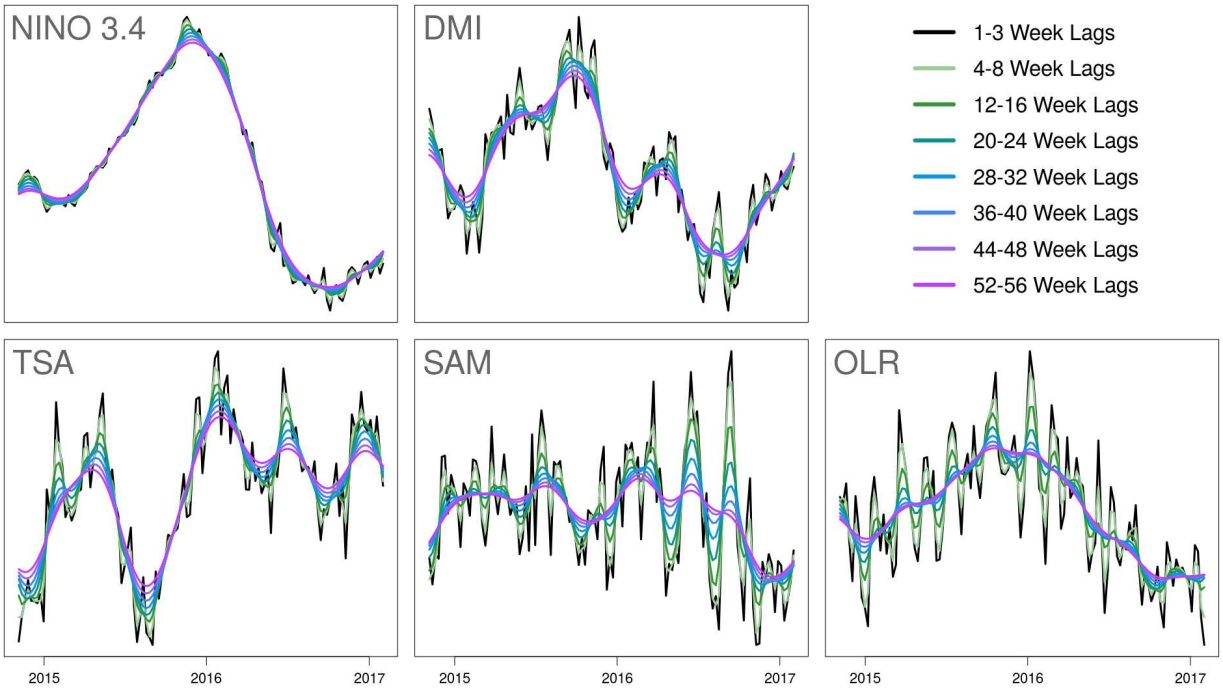

Figure 2.4 Black curve shows the original climate index data, which is used for lags of one through three weeks. Colored curves show every other level of smoothing applied to the climate index data, which is used for lags of four through 52 weeks. Vertical axis has been omitted for visual clarity.

2.4 Variable Selection and Model Fitting

We consider 52 lags of each climate mode index, as well as quadratic terms and pairwise interactions. This results in far more covariates than observations. We want to perform some form of variable and lag selection. Buchholz *et al.* (2018) broke this process up into two parts. First, they iterated through all possible lag combinations. At a given combination of lag values (called a "lag set"), each index was fixed at a single lag value. Stepwise selection was then used for variable selection. This resulted in a list of optimally performing models (one model for each lag set). From this list, a single model was selected based on adjusted R^2 to represent the climate-CO relationship for the given response region.

By iterating through the lag sets, Buchholz *et al.* (2018) was able to use stepwise variable selection without the need for large computational resources. This is because for a fixed set of lag values, the number of possible covariates is relatively small. The number of possible covariates would be much larger if instead all possible lags for each index were considered simultaneously, making stepwise selection impractical based on computational requirements.

Here we use regularization for both variable and lag selection. Because regularization is well suited for cases with more covariates than observations, we are able to consider all possible lag values for each index simultaneously, without requiring large computational resources. A general expression for the coefficient estimates generated by regularization is given by

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \sum_{j=1}^{p} p(\beta_j), \tag{2.2}$$

where β is a vector containing all coefficients (a_k, b_{ij}, and c_l in Equation 2.1) corresponding to the covariates in X (χ_k, $\chi_i \cdot \chi_j$, and χ_l^2 in Equation 2.1), Y is the response, and $p(\beta)$ is some penalty applied to the coefficients. In Equation 2.2, i iterates through the number of observations and j iterates through the number of covariates. The first term is the sum of squared residuals and can be thought of as a measure of fit. The LASSO penalty, given by

$$p(\beta) = \lambda|\beta| \tag{2.3}$$

has the added benefit of shrinking coefficient estimates to exactly zero, hence performing variable selection (and lag selection for our application). The tuning parameter, $\lambda \geq 0$, is a free parameter that balances the fit term and the penalty term. We discuss our method for selecting λ values shortly.

Instead of the traditional 1-norm used in the LASSO, we apply a slightly more flexible penalty: the minimax concave penalty (MCP). The MCP penalty is given by

$$p(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\eta} & \text{if } |\beta| \leq \eta\lambda \\ \frac{\eta\lambda^2}{2} & \text{otherwise.} \end{cases} \tag{2.4}$$

While the LASSO penalty increases linearly with $|\beta|$, the MCP penalty gradually levels off until eventually applying a constant penalty after $|\beta|$ surpasses a threshold defined by the free parameter $\eta \geq 1$. Again, we discuss our method for selecting η values shortly. The MCP results in less biased estimates for non-zero regression coefficients (Zhang, 2010). Essentially, it allows for larger coefficient estimates on the significant terms (which might be closer to the "true" model). We found that using the MCP penalty over the 1-norm penalty from the LASSO increased model performance. The price we pay for this generality is the introduction of a second parameter, η, in additional to the traditional tuning parameter, λ, that weights the penalty term.

The typical regularization fitting procedure involves minimizing the loss function (i.e. Equation 2.2) for a sequence of λ values, called a solution path. A single model is then selected from the solution path using an information criterion (e.g. AIC or BIC) or cross-validation test error. Here we use a more general form of the BIC, called the Extended Bayesian Information Criterion (EBIC), given by

$$BIC_\gamma(s) = BIC(s) + 2\gamma \log \tau(s), \tag{2.5}$$

where s is the model being evaluated, BIC is the standard form of the BIC, τ is the number of possible models with equation dimension (i.e. number of terms) as s, and $\gamma \in [0, 1]$ controls the extra penalty contained in the second term.

The EBIC can apply a much stronger penalty to large models (i.e. models with many selected terms) than the BIC. This is well suited for applications in which the number of possible covariates is large, but the true model might in fact be quite small. Since we believe this to be the case for the atmospheric CO application, we use the EBIC rather than the BIC or cross-validation test error to select λ.

With these more flexible adaptations to the traditional LASSO, we are left with a number of free parameters: λ, the tuning parameter, η, which controls the MCP penalty, and γ, which controls the EBIC. For a given combination of these parameters, we fit the coefficients using the `RAMP` package in `R` (Hao *et al.*, 2018). `RAMP` is a recent regularization method that efficiently computes a hierarchy-preserving solution path for quadratic regression (i.e. models including squared and interaction terms). Enforcing hierarchy, or more specifically strong hierarchy, requires that terms present in an interaction are also present as main effects. Strong hierarchy (also known as the marginality principle) has long been recommended for models with interactions, as it helps avoid misinterpretation of the included covariates (Nelder, 1977). Another benefit of the `RAMP` algorithm is its remarkable efficiency. `RAMP` is able to compute full solution paths much faster than similar hierarchy-preserving algorithms available in `R`, such as `hierNet` (Bien *et al.*, 2013) or `ncvreg` (Breheny & Huang, 2011).

We select parameter values with a simple grid search broken into two steps:

1. Select a γ value on $[0, 1]$. Values closer to 0 will result in larger models and values closer to 1 will result in smaller models. We discuss γ value selection in more detail in the following sections.

2. For the given γ value, vary λ and η simultaneously. For each combination of λ and η, fit regression coefficients using the `RAMP` package. Select the model that minimizes the EBIC computed with the selected γ value.

    - The `RAMP` algorithm automatically computes a data-driven sequence of λ values, so no user input is required.
    - We vary η on a logarithmic sequence from 1.001 to 6. This range was selected manually by trial-and-error and tuned specifically for this application. We tested this range on a number of different covariate combinations and response regions (including MSEA), and the selected η value always fell well within this range. Note that the optimal η value is completely data dependent and this sequence will need to be adjusted for different applications or data.

This grid search results in an optimal model, where optimal is defined by the choice of γ. In the remainder of this chapter, we discuss how this modeling framework can be used to address our two goals of interpretability and prediction.

2.5 Research Focus #1: Interpretable Models for Scientific Conclusions

Here we focus on interpreting the selected models, rather than their predictions. In addition to being useful tools for prediction, these models can help explain the connections between climate and atmospheric chemistry. For instance, scientists might be interested in the question: which index (and hence which aspect of the global climate system) has the most influence on atmospheric CO loading in Maritime Southeast Asia? Answering this will, for instance, help make predictions of the global carbon budget. We can provide an elementary answer to this question by examining the selected model terms and lags.

2.5.1 Framework for Identifying Optimally Performing Models at Various Complexities

By simply varying γ over a range of values on $[0, 1]$, we can create a list of "optimally performing" models. Optimal here refers to the fact that these models are the result of a grid search over the other two free parameters, λ and η. The resulting models decrease in complexity (i.e. number of terms), as larger γ values make the EBIC's penalty on large models stronger. Note that after selecting model terms via the `RAMP` algorithm, we refit the coefficient estimates with their maximum likelihood estimates.

For the MSEA region, this procedure results in the models listed in Figure 2.5. The color of each box corresponds to the γ value that was used to generate it. A number of γ values produce the same models. In each box, the name of the index and the corresponding lag is listed (in the format "name_lag"), along with the coefficient estimates and standard error.

Moving from left to right in Figure 2.5, we see that the models decrease in size (from 17 terms to nine), while their performance drops only slightly (from explaining 70% of variability in the response to 61%). By examining the terms that remain in these models as they become more parsimonious, we can determine which indices and lags are most influential in explaining variability in the response.

For the MSEA region, we can see that the Nino 3.4 index lagged at four weeks remains in all of the models with a positive coefficient estimate. This makes sense, as ENSO is a major climate driver in the tropics, with positive anomalies resulting in warmer, drier conditions (Nur'utami & Hidayat, 2016). The lag of four weeks indicates that it takes about four weeks for the effect of a Nino 3.4 anomaly to impact CO anomalies. Additionally, the Nino 3.4 lag of four weeks appears as a squared term in the most parsimonious model, indicating that there is a nonlinear relationship between Nino 3.4 and CO. This is confirmed by examining the residuals of a model fit to solely the Nino 3.4 lag of four weeks (not shown).

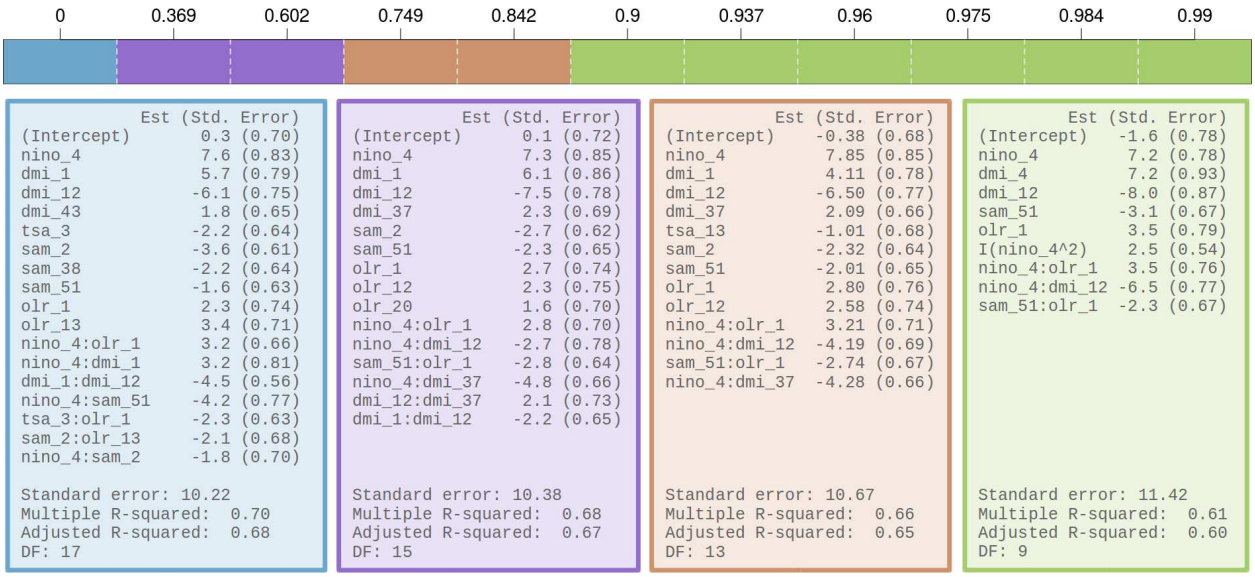

Figure 2.5 Optimal models for the MSEA region for a sequence of γ values. Note that multiple γ values often produce the same model. The color of each box corresponds to the γ value that was used to generate the model contained within it. Within each box, the selected model terms are listed in the format "name_lag," where lags are in weeks. Coefficient estimates and standard errors are listed for each term, and summary statistics are listed below each model. Note that "Nino" refers to the Nino 3.4 index.

The selected DMI lags also suggest an interesting relationship. A DMI lag of 12 weeks remains in the models as they become more parsimonious, as well as a shorter lag (which switches from one to four weeks between the smallest two models). The sign of the larger lag is negative, while the sign of the shorter lag is positive. Positive DMI anomalies are associated with reduced rainfall in parts of MSEA, while negative DMI anomalies are associated with increased rainfall (Nur'utami & Hidayat, 2016).

One possible explanation for the selection of these lags is as follows. A negative DMI anomaly causes increased rainfall in parts of MSEA. As a result of the rainfall, there is increased vegetation growth over the next two months (and hence an increase in the amount of available biomass for fuel). A subsequent positive DMI anomaly causes decreased rainfall. As a result, the biomass that has accumulated over the last two months dries out, making it more prone to fire and hence increased CO loading. Note that the DMI lag of 12 weeks might be included solely to allow for an interaction with the Nino 3.4 lag of four weeks. However, given the large coefficient on the DMI lag of 12 weeks, we believe that this main effect is significant.

A positive OLR lagged at one week remains in the MSEA model as it becomes more parsimonious. This again makes sense, as positive OLR anomalies are associated with less cloud cover and hence less rain. Therefore, we can infer that a decrease in rain increases the probability of fire and hence CO loading in the short term. The TSA index, on the other hand, is only included in the largest model. This could be because the TSA describes sea surface temperatures in the southern Atlantic Ocean, which is very far from

the MSEA response region. Therefore, it makes sense that the TSA is less important than the other indices in explaining CO variability in MSEA, as the other indices are based on aspects of the global climate system located closer to MSEA.

Finally, two Nino 3.4 interaction terms remain in the model as it becomes more parsimonious. One interaction is with the OLR at a one week lag and the other is with the DMI at a 12 week lag. The sign of these interaction terms is the same as the non-Nino 3.4 component. This could indicate that the effects of these indices are amplified when they are in phase, a result that has been previously identified in the literature (Cleverly *et al.*, 2016; Nur'utami & Hidayat, 2016).

2.5.2 Assessing Stability of Selected Model Terms

While the scientific conclusions drawn in the previous section are interesting and seem to broadly agree with literature, we want to ensure that the selected covariates are in fact meaningful. That is, we want to avoid over interpreting the role of covariates if slight changes in data result in drastically different models, as these models would not be capturing a meaningful physics-based relationship but would rather be artifacts of the specific training data.

Therefore, we perform one-year-out resampling to assess the "stability" of the selected models. More precisely, we do the following.

1. Iterate through the years present in the data. For this application, this spans 2001 to 2019.
2. For each year, create two data sets:
    - Training set: This set consists of all data except for the given year.
    - Testing set: This set consists of only the given year left out of the training set.
3. Using the training set, refit two different models:
    - Main model: This model maintains the same form as the model trained on all of the data. That is, we force it to retain the same covariates. However, we refit the coefficients to the training set.
    - New model: This model we completely recreate based on the training set. We do not force it to take the same form as the model trained on all of the data.
4. Make predictions of the testing set using both the main model and the new model. Compute the root mean square error (RMSE) of both sets of predictions.

Note that we perform this resampling on the largest model from Figure 2.5 (i.e. the model with the most terms) because it contains most of the terms present in the more parsimonious models as well as extra terms that result in slightly higher predictive skill.

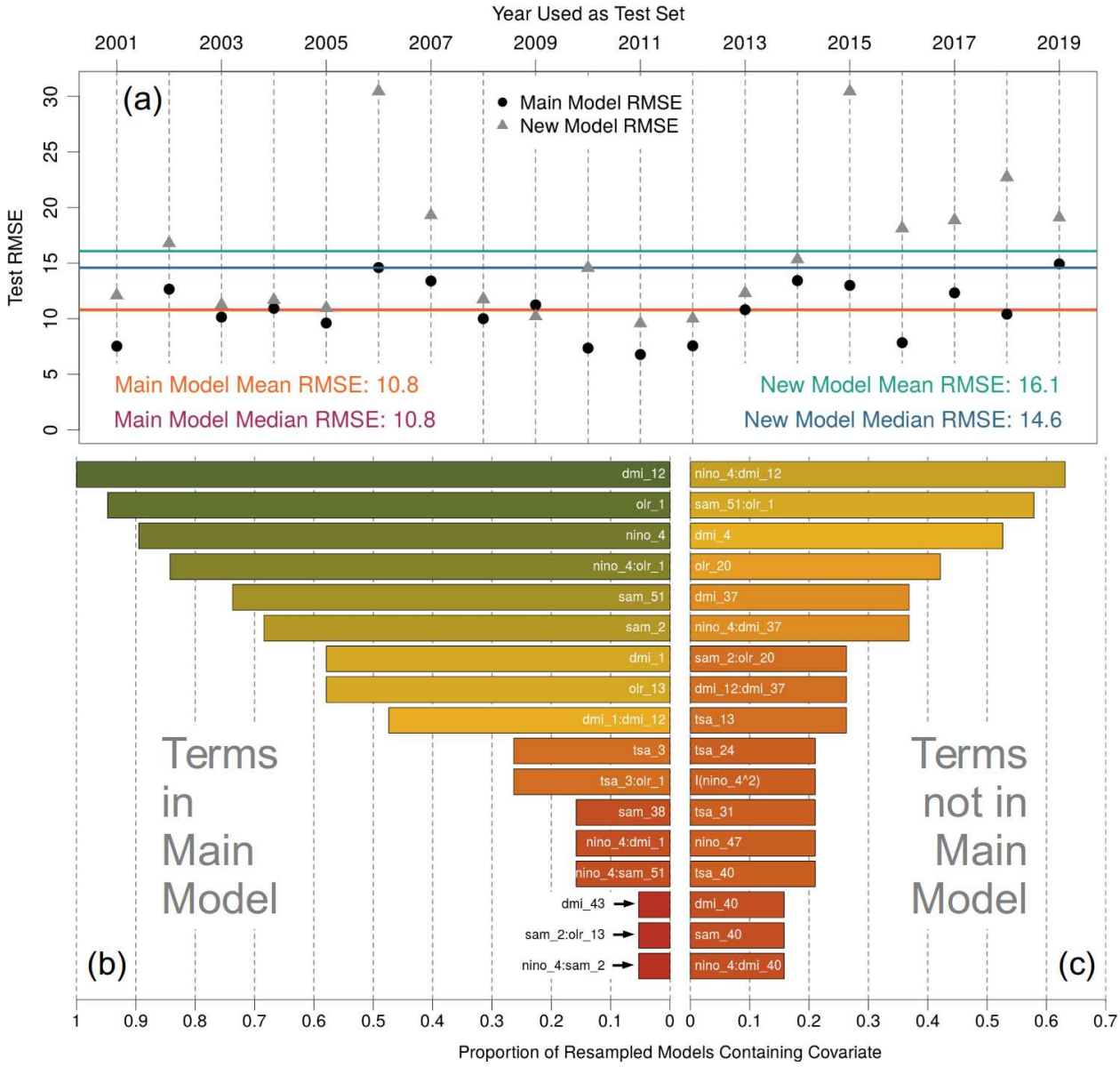

Figure 2.6 Results from the one-year-out resampling. Main model refers to the model forced to retain the structure of the model trained on all of the data, but with refit coefficient estimates. New model refers to the model allowed to completely change according to the particular training set. (a) shows the out of sample prediction error for each training set. The year on the horizontal axis indicates which year was used to test the models. The main model almost always out performs the new model. (b) shows the frequency with which main model terms appear in the new models. Similarly (c) shows the frequency with which terms not present in the main model appear in the new model. We see that the most significant terms appear in many of the retrained models. The color in (b) and (c) corresponds to the proportion on the horizontal axis and is included for visual clarity. Note that "Nino" refers to the Nino 3.4 index.

Figure 2.6 shows the results of this resampling and is divided into three sections. Figure 2.6(a) shows the RMSE of the predictions described above. The RMSE of the main model (that is, the model that retains the structure of the model trained on all data) tends to perform as well or better than the model

allowed to completely change according to the new training set. This provides justification for using the main model to predict out of sample CO anomalies (i.e. future CO anomalies using current climate data). Note that the RMSE of the new model is significantly larger when 2006 and 2015 are left out of the training set. These years have some of the largest CO anomalies (see Figure 2.2), which might indicate that these extreme years are important in driving the form of the model.

Figure 2.6(b) and Figure 2.6(c) show how often certain model terms appear in the new models (that is, the models allowed to completely change according to the new training data). This gives some indication of the stability of the various model terms. If a term is present in many of the retrained models, then the modeling framework is likely picking up a physics-based relationship. Terms that are absent from many of the retrained models are more likely artifacts of the specific training set, rather than a true physical relationship.

Figure 2.6(b) shows how often the main model terms reappear in the new models. Notably, the terms present in the most parsimonious model from Figure 2.5 are most likely to appear in the retrained models. This indicates that these terms are explaining the most stable aspect of the physical relationship. Other terms, such as the 43 week DMI lag, rarely appear in the retrained models. This indicates that less consideration should be given to these terms when attempting to explain the physical relationship between climate and CO.

Figure 2.6(c) shows how often terms not present in the main model appear in the retrained models. Note the different scales on the horizontal axis between subfigures (b) and (c). Here we see that a selection of terms not in the main model appear relatively frequently in the retrained models. Recall that when moving from the second smallest to the smallest model in Figure 2.5, the shorter DMI lag switches from one week to four weeks. In Figure 2.6, we see that both the one and four week DMI lags show up in about half of the retrained models. This indicates that these terms are interchangeable, and determining which is included likely depends on the other selected covariates.

Figure 2.6(b) and Figure 2.6(c) further confirm that the terms present in the most parsimonious model for the region (see Figure 2.5) are capturing meaningful signal and are not simply artifacts of the specific training set. This is because they remain in a large majority of the retrained models, which iteratively remove one year of data from the training set. Furthermore, Figure 2.6(c) illustrates that the interaction between Nino 3.4 lagged at four weeks and DMI lagged at 12 weeks, although not present in the main model, is still a significant interaction in explaining CO variability in MSEA. This also holds for the interaction between SAM lagged at 51 weeks and OLR lagged at one week. The terms that are included less often in the retrained models are likely more data dependent and help the model capture subtleties in the response. As a result, it is more likely that these terms would change with small changes in the data.

An example is the TSA term lagged at three weeks present in the main model. This term appears in less than 30% of the retrained models, which confirms the analysis in Section 2.5.1 that finds that TSA is less important in explaining CO variability in MSEA.

This analysis is useful when interpreting the selected model terms and using them to draw scientific conclusions. Figure 2.6 provides justification for assigning scientific weight to the terms that remain in the most parsimonious model in Figure 2.5.

2.6 Research Focus #2: Models with Predictive Skill

We now turn our attention to the predictive skill of the selected models, rather than their form or the selected indices and lag values. There is obvious value in making advanced predictions of atmospheric CO loading (again, a proxy for fire intensity, especially in the Southern Hemisphere). Advanced warning of a particularly intense fire season would give governments enough time to properly staff fire departments, stock up on masks, and warn citizens in high risk areas.

2.6.1 Model Predictions

The largest model from Figure 2.5 is used in this section, as we now focus on predictive skill rather than model interpretability. In Figure 2.7 we show weekly observations and fitted values from two model variants. Note that these predictions are in-sample, meaning that we are showing predictions of the observations used to train the model. In Figure 2.7(a) we show predictions for the entire time series, with summary statistics shown in the top right corners. The top plot shows predictions from a model without the OLR index, while the bottom plot shows predictions from the full model (i.e. the model presented in Figure 2.5). We can see that adding the OLR results in a slight increase in R^2 and a slight decrease in RMSE.

Furthermore, in Figure 2.7(b) and (c), we highlight two of the most anomalous years, which shows that the OLR helps capture the extreme CO anomalies. For 2015 in particular, this makes sense, as the MJO experienced an extreme anomaly during this time (which we attempt to capture with OLR).

In Figure 2.8(a) we show month-averaged observations and predictions from two different model variants for the entire time series. The top plot shows predictions from a month-based model. To create this model, we took month-averages of the predictor variables and then trained the model on only these month-averaged covariates using the framework presented in Section 2.4. The bottom plot instead shows the month-averaged predictions from the model trained on the weekly data (i.e. the model shown in Figure 2.5). We see a noticeable increase in model performance when using the weekly data, suggesting that the weekly data is able to capture meaningful signal beyond the month-averages.

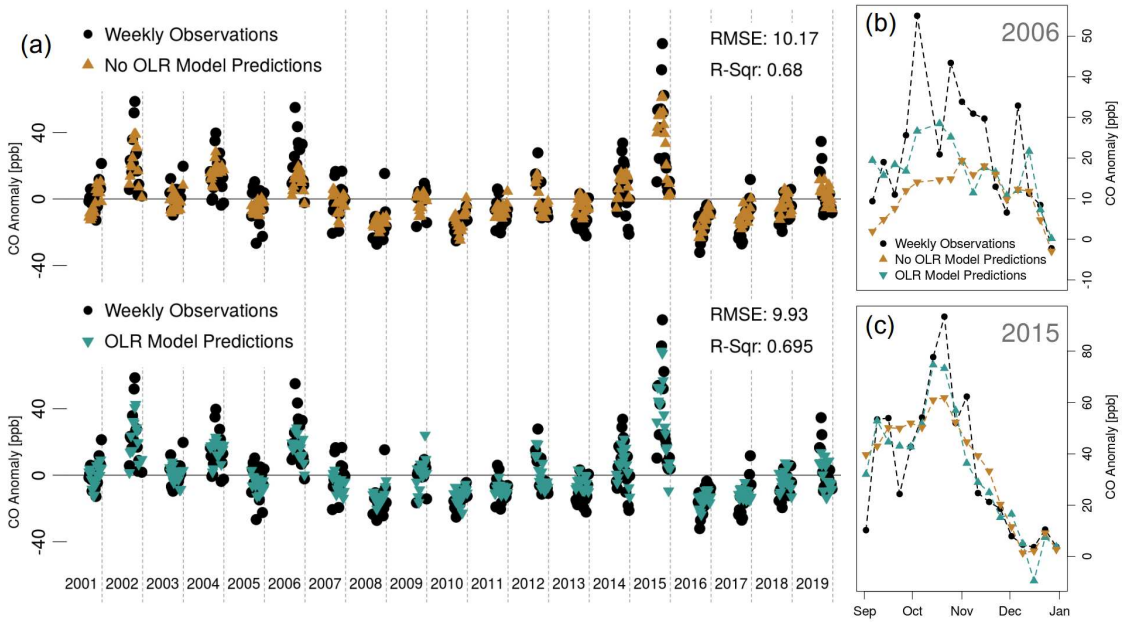

Figure 2.7 In-sample predictions from two model variants. In (a), the top plot shows predictions from the optimal model without the OLR, and the bottom plot shows predictions from the optimal model with the OLR. Adding the OLR appears to increase predictive skill during the extreme CO anomalies shown in (b) and (c).

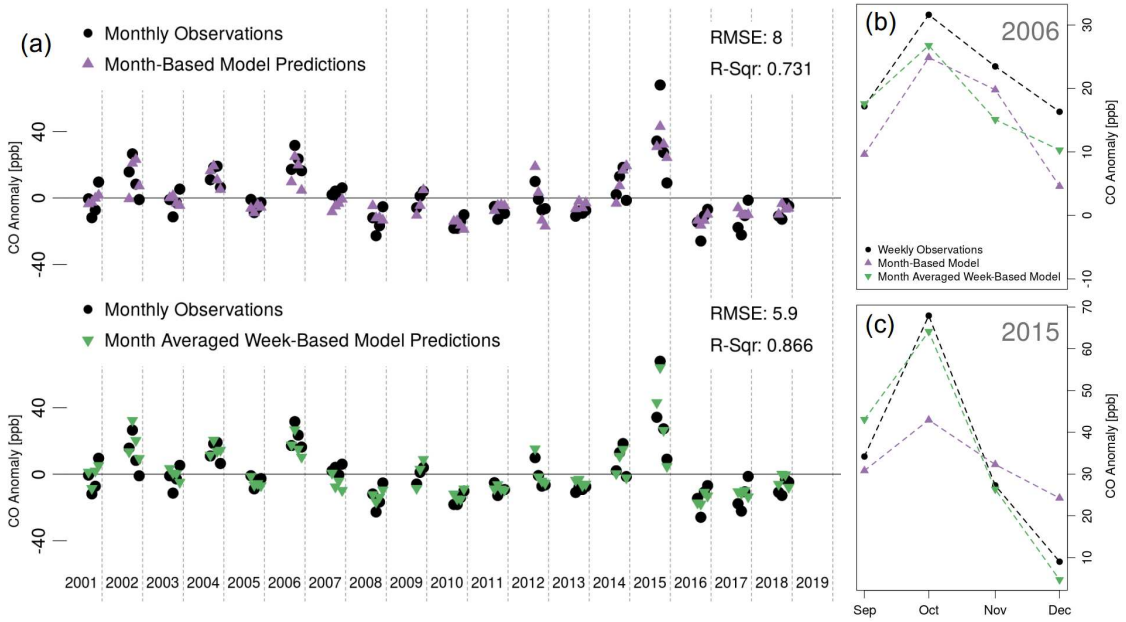

Figure 2.8 In-sample predictions from two additional model variants. In (a), the top plot shows predictions from a model trained on month-averaged covariates, and the bottom plot shows month-averaged predictions from a model trained on week-averaged covariates. The increase in model performance indicates that there is meaningful signal in the higher frequency climate index data, which is clearly seen in the anomalous years shown in (b) and (c).

This is an interesting result, as it suggests that the higher frequency signals present in the climate indices are in fact meaningful signal and not simply noise. This is perhaps most important for OLR (the proxy for localized MJO), which has a higher frequency component than the other included climate indices. This increase in performance can be seen clearly during the 2015 CO anomaly.

Note that Buchholz *et al.* (2018) present models based on month-averaged data. Here we wanted to confirm that there was meaningful signal in the week-averaged data (i.e. that the weekly signal was not just the monthly signal with additional noise). Figure 2.8 shows that there is in fact meaningful signal in the weekly data, making these models with a higher temporal resolution an improvement over those presented in Buchholz *et al.* (2018).

2.6.2 Increasing Minimum Lag Threshold

The model predictions shown in the previous section are useful for demonstrating model performance and the comparative benefit of using the OLR and week-averaged data. However, these models include an OLR term lagged at one week, which significantly reduces their practical utility. This model can only predict as far in advance as the length of its smallest lag, or in this case, one week. Predictions with longer lead times would give governments more time to prepare for intense fire seasons.

To increase the prediction horizon, we implement a minimum lag threshold to the modeling framework that only allows lags greater than the threshold value to be selected. Because increasing this threshold reduces the number of possible covariates, we also extend the maximum lag value as the minimum lag threshold is increased. Specifically, we consider lags between the minimum lag threshold and 52 weeks plus this threshold. This ensures that all models are based on one year of climate data, making it easier to compare their predictive skill.

Figure 2.9 shows a selection of model performance metrics as this minimum lag threshold is increased. Note that this figure is for the MSEA response region and considers the largest model from the range of EBIC γ values (i.e. $\gamma = 0$).

The top plot in Figure 2.5 shows the number of terms in the selected model for each minimum lag threshold. We can see that the models tend to have around 17 terms, although some have less. This will have a slight effect on the performance metrics, albeit a small one. The second plot shows the R^2 value of the selected models. As expected, the model performance drops off as the minimum lag is increased. However, this decline is not very rapid. That is, models with a high minimum lag threshold still explain a large percent of the variability in atmospheric CO anomalies. This is promising, as it means that predictions can be made farther in advance without losing too much predictive skill.

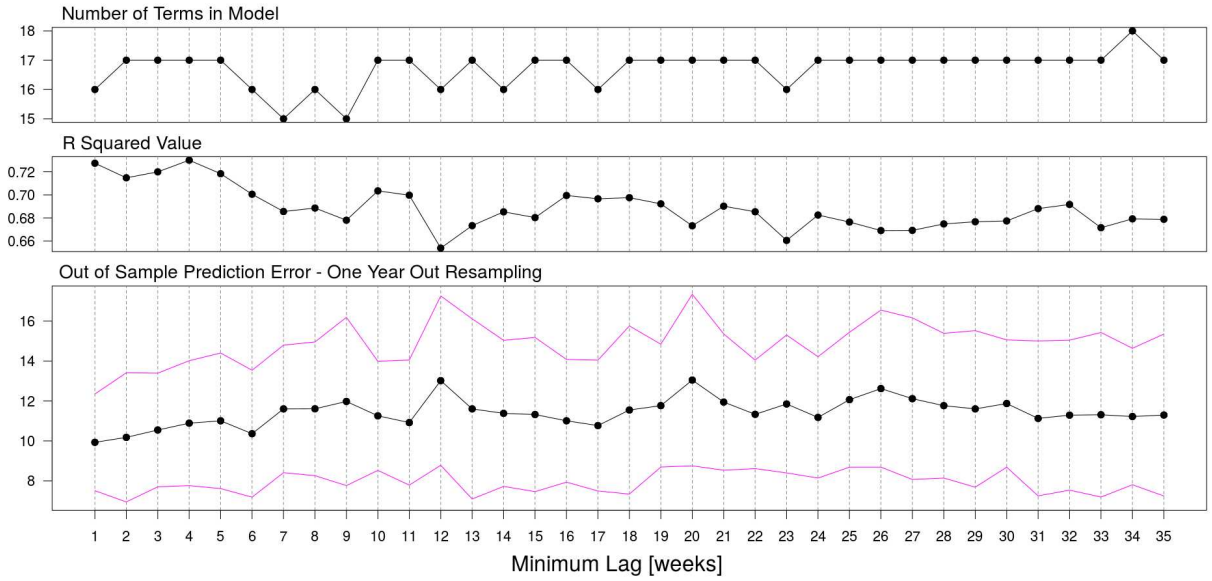

Figure 2.9 Model performance for the MSEA response region at increasing minimum lag thresholds. Top plot shows the number of terms in the selected model. Middle plot shows the R^2 value of the selected model. Bottom plot shows an average out of sample prediction error for each model with magenta lines showing $\pm$ one standard deviation. Here we iteratively leave one year out, train the model on the remaining data, and test it on the left out year. Plotted is the average RMSE with $\pm$ one standard deviation lines in magenta from this procedure as a function of minimum lag. We can see that model performance drops off with an increasing minimum lag threshold, although at a fairly gradual pace.

The third plot shows another performance metric: the average out of sample prediction error from a similar one-year-out resampling study. Here we successively leave one year out, train the model on the remaining data, and test it on the left out year. The average RMSE is then taken for each different training and testing set pair and plotted as a function of minimum lag threshold. We again see that performance falls off, although gradually.

We believe that the gradual decline of model performance is due to the highly auto-correlated nature of the climate indices used here as predictor variables (not shown). Since many of the short lags are highly correlated to larger lags of the same index, we believe that these larger lags are able to explain much of the same CO variability when the shorter lags are excluded. This is again promising, as it means that predictions can be made decently far in advance (on the order of a half year) without dramatically compromising performance.

To further visualize model performance at increasingly large minimum lag thresholds, we consider predictions for the 2015 CO event in the MSEA region. Figure 2.10 shows predictions from the models corresponding to the minimum lag thresholds from Figure 2.9. In Figure 2.10, the color scale represents the CO anomaly, and the horizontal axis represents time. The bar along the bottom of the figure shows CO

observations from 2015. The remainder of the vertical axis corresponds to the minimum lag threshold used to fit the models, and hence each row of the figure corresponds to predictions from a different model. The predictions largely capture the structure of the CO observations for minimum lag thresholds below 25 weeks (about six months). After this point, the predictions begin to flatten out (i.e. not capture the extremes in the response) and the predicted spike drifts earlier in the year (starting around late September instead of mid October). This result largely agrees with Shawki *et al.* (2017), who found that a drought metric could be reasonably predicted 180 days (about 25 weeks) in advance. However, unlike Shawki *et al.* (2017), our predictions rely solely on past climate mode index anomalies, rather than forecasts from a global climate model.

We therefore believe that these models can be useful for predicting the structure of the CO anomalies up to six months in advance for MSEA. However, if a very high level of fidelity is required on a weekly timescale, then restricting predictions to less than a three month lead time is advised.

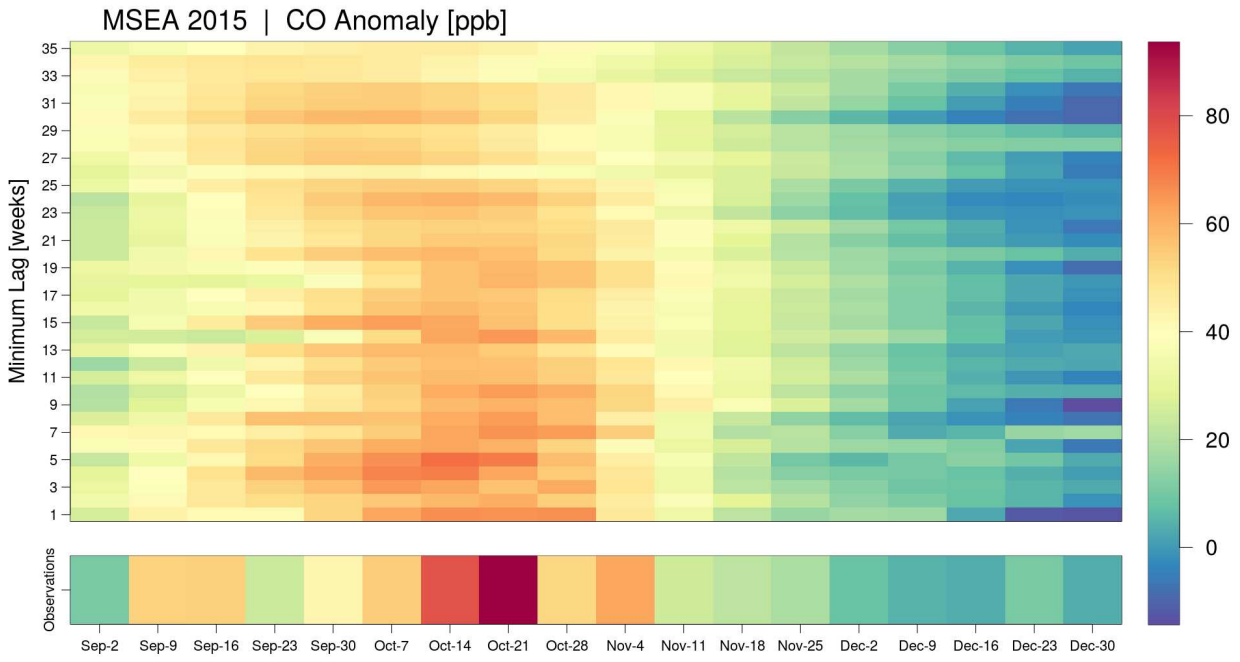

Figure 2.10 Predictions of the 2015 CO anomalies in the MSEA response region for a range of minimum lag thresholds. Color represents the CO anomalies, and the horizontal axis represents time. Observations are shown as a horizontal bar along the bottom of the figure. The remaining vertical axis corresponds to the minimum lag value, and hence each row of the figure is a prediction from a different model. We see that the general structure of the observed CO anomalies is preserved for minimum lags under 25 weeks (about half a year).

2.7 Summary

Here we build on previous work aimed at explaining the relationship between climate and atmospheric CO variability. Again, atmospheric CO is a useful proxy for fire intensity in the Southern Hemisphere, as fires are the main source of CO variability in the Southern Hemisphere and CO is easy to remotely sense on a global scale.

We have developed a regularization framework that highlights a variety of optimally performing models at decreasing complexities, isolating the most important indices and lag values as the models become more parsimonious. Notably, for the MSEA response region, we identify the Nino 3.4 index lagged at four weeks as a primary driver of atmospheric CO. Other important climate indices are the DMI and OLR (as a proxy for the MJO). We further identify that Nino 3.4 interactions with the OLR and DMI are significant predictors, suggesting that the effect of these indices is amplified when they are in phase.

Note that these findings largely agree and expand upon the results in Buchholz *et al.* (2018). For the MSEA region, Buchholz *et al.* (2018) found that a Nino 3.4 lag of one month, DMI lag of eight months, TSA lag of five months, and SAM lag of one month were important predictors. The largest model presented in this study contains a Nino 3.4 lag of four weeks, DMI lag of 43 weeks, TSA lag of three weeks, and SAM lag of two weeks. All but the TSA term (which we have shown is not very important for the MSEA region) agree closely on their selected lag. However, the models we present here are capable of including multiple lags of a single index, which expands on the work in Buchholz *et al.* (2018) and highlights more complex relationships between climate and CO.

We also performed a sensitivity analysis in which we use leave one-year-out resampling to quantify the robustness of the "main model" fit to all of the data, justifying its use in predicting future CO anomalies. Additionally, we determine which model terms are most likely to remain in the model with slightly different data, finding that the model terms that remain in the most parsimonious model are very likely to remain in the resampled models. This justifies assigning scientific weight to the selection of these model terms, as it suggests that their inclusion in the model is not an artifact of the specific training data used.

Furthermore, in-sample model predictions for the MSEA response region can explain around 70% of the variability in the weekly CO anomalies solely using climate indices as predictor variables. We further use these predictions to highlight the importance of the OLR (as a proxy for the MJO) in overall model performance and in capturing the most extreme CO anomalies. Similarly, we show that month-averaged predictions from a model trained on week-averaged data outperform predictions from a model trained on month-averaged data. This suggests that there is meaningful signal in the week-averaged data and justifies its use over month-averaged data.

These predictions present an improvement over the models in Buchholz *et al.* (2018). When using week-averaged data to train the model, we are able to explain 87% of the variability in the month-averaged CO observations. The model in Buchholz *et al.* (2018) explains 75% of the month-averaged CO. This increase in predictive skill is likely a result of: 1) the ability to include multiple lags of a single climate mode index, 2) the additional signal contained in the week-averaged data, and 3) the inclusion of the OLR proxy index.

Finally, we perform a minimum lag threshold study to push the predictive capabilities of these models. We find that models for the MSEA response region are still able to explain around 65% of the weekly atmospheric CO variability when forced to only use lags greater than 35 weeks. This indicates that predictions can be made relatively far in advance without losing the overall structure and general amplitude of the CO anomalies. If these models are to provide advanced warning of fire season intensity, then this is beneficial because it extends the time available to prepare.

CHAPTER 3
A HIERARCHICAL SPATIAL MODEL FOR ESTIMATING METHANE FIELDS FROM REMOTELY SENSED OBSERVATIONS WITH NOISE

3.1 Introduction

Natural gas is often referred to as a "bridge fuel" between other fossil fuels and renewable energy. Methane (the primary component of natural gas) produces less carbon than both coal and oil when combusted and has a shorter lifetime than carbon dioxide if released into the atmosphere (EIA, 2020; EPA, 2020). However, methane absorbs more energy than carbon dioxide, making it a much more potent greenhouse gas (EPA, 2020). Therefore, if methane is to be considered a cleaner alternative to other fossil fuels, its production, transportation, and storage must be done in a way that limits both fugitive emissions (i.e. leaks) and operational emissions (i.e. venting). To ensure that this is being done, effective emissions monitoring is required.

Emissions monitoring can be performed using an array of platforms: ground-based "fenceline" sensors, aircraft, and satellites. The use of fenceline sensors is relatively new, but may improve monitoring because of their high sampling rate and positioning close to the oil and gas facilities. However, they are hard to deploy across an entire basin because of their operational and deployment costs, making large scale emission estimates challenging (Fox *et al.*, 2019). Aircraft can be used to provide top-down emission estimates over a larger spatial domain than individual fenceline sensors, but these overflight campaigns are costly and are often only performed in response to a previously identified leak (Hirst *et al.*, 2013).

Compared to the other monitoring platforms, satellites provide the largest spatial coverage, often spanning the entire globe on a daily basis. This makes them an important tool when quantifying the global carbon budget (Crowell *et al.*, 2019). However, satellite observations are limited by their coarse spatial resolution, making it very challenging to isolate specific emission sources. The highest resolution methane data made publicly available has a resolution of 7×5.5 km, meaning that specific sources within this area cannot be distinguished (Hu *et al.*, 2016). Commercial options, such as the Montreal-based GHGSat, provide much higher resolution data (on the order of 50×50 m) by focusing on a single region, rather than providing global coverage (Jacob *et al.*, 2016). However, these data are not publicly available.

Here we focus on methane observations from The Tropospheric Monitoring Instrument (TROPOMI) on board the Sentinel-5 satellite (Veefkind *et al.*, 2012), as satellite-based observations have great potential for supplementing other methane monitoring platforms despite their relatively coarse resolution. Much research has gone into estimating basin-wide or global emissions using satellite data. These estimates are

often the result of inversion studies that utilize transport models. Zhang *et al.* (2020) use TROPOMI data and a nested version of the GEOS-Chem transport model to estimate methane emissions from the Permian basin in the United States. Similarly, Crowell *et al.* (2019) compile an ensemble of flux inversion models that use data from the OCO-2 instrument to estimate global carbon dioxide fluxes. Other notable studies include Kort *et al.* (2014), Turner *et al.* (2015), and Buchwitz *et al.* (2017). These estimates are extremely important when compiling the global carbon budget and provide a useful comparison to bottom-up engineering or inventory estimates, often finding that the bottom-up estimates undercount total emissions (Turner *et al.*, 2015; Zhang *et al.*, 2020).

These large basin-wide studies, however, do not focus on isolating specific emission sources. Other work has focused on this task, such as Varon *et al.* (2019), which identifies plume structure in the TROPOMI data and uses source rate retrieval algorithms to estimate emission fluxes. However, these rate retrieval algorithms require that the methane plume spans multiple TROPOMI pixels, making it challenging to isolate small to medium sized leaks.

Here we present a framework for predicting methane concentrations at sub-pixel resolution. The purpose of this work is to localize small scale emissions that would not otherwise be detectable (i.e. that do not span multiple TROPOMI pixels). This could be useful for rapid event identification on a small scale and for comparison to the other emissions monitoring techniques described above. Specifically, we seek to answer the following questions:

1. Can we say something meaningful about the continuous methane field on a scale smaller than TROPOMI's footprint?

2. Can we quantify the uncertainty in our estimate of this continuous field?

To address these questions, we have developed a hierarchical spatial model to estimate the methane field given a set of TROPOMI observations and their corresponding footprints, which builds on previous work. The problem of estimating a hidden or unknown field given noisy observations of that field is well studied. This is known as geostatistical prediction, Kriging, or optimum interpolation (e.g. Cressie (1993)). Two issues are often present when Kriging satellite data. The first is the need to model non-stationary covariance structures, which is not accounted for in the traditional Kriging spatial model. The second is a computational bottleneck that results from estimating statistical parameters for the covariance function. This process scales as $\mathcal{O}(n^3)$ operations, where n is the number of observations, which quickly becomes infeasible for large n. Much work has gone into addressing these issues, such as Nychka *et al.* (2002) and Cressie & Johannesson (2008).

Instead of addressing these issues, however, we instead focus on how TROPOMI's footprint affects the Kriging estimate. Since this work focuses on the localization of small scale emissions, we can work with relatively small spatial domains (i.e. a geological basin on the scale of 54 square miles). This makes the non-stationarity and computational concerns typically encountered when Kriging satellite data less of an issue.

We expand on the general Kriging estimate to account for the shape of the TROPOMI footprint, broadly following the structure of the data model presented in Nguyen *et al.* (2017). This model assumes the observations from the satellite instrument are spatial averages over the unknown field within each footprint plus a noise term. While Nguyen *et al.* (2017) focus on using the model for spatial data fusion between satellite instruments, we focus instead on sub-footprint predictions, with the ultimate goal of monitoring or localizing small scale methane emissions from the oil and gas industry. Work in this direction can be used to fuse monitoring data across platforms (i.e. between fenceline sensors, aircraft, and satellite).

The rest of this chapter is organized as follows. In Section 3.2, we describe the TROPOMI data used in this study. In Section 3.3, we outline the hierarchical spatial model we have created to estimate the methane field from the TROPOMI observations, and in Section 3.4 we discuss parameter estimation for this model. In Section 3.5 we discuss our methods for making predictions and quantifying their uncertainty, and in Section 3.6 we apply this model to a single TROPOMI overpass of northeast Colorado. Finally, we discuss our results and future work in Section 3.7.

3.2 Methane Observations from the TROPOMI Instrument

The Tropospheric Monitoring Instrument (TROPOMI) is the scientific instrument on board the Copernicus Sentinel-5 Precursor satellite. TROPOMI was launched on October 13, 2017 and has been providing publicly available data since April 30, 2018. TROPOMI observes each geographical point approximately once per day with a spatial resolution of 7×5.5 km as of August 2019. Earlier observations had a resolution of 7×7 km (Veefkind *et al.*, 2012). The resolution of the TROPOMI instrument provides a remarkable improvement over similar scientific instruments. The Ozone Monitoring Instrument (OMI) on board the Aura satellite has a resolution of 13×24 km and the Global Ozone Monitoring Experiment-2 (GOME-2) on board the METOP-A satellite has a resolution of 80×40 km (Levelt *et al.*, 2006; Munro *et al.*, 2016). TROPOMI retrievals include a number of trace gasses, including ozone, carbon monoxide, nitrogen dioxide, sulfur dioxide, and methane. Here we focus on methane.

The two largest anthropogenic sources of methane in the United States are natural gas production and enteric fermentation (i.e. the digestive process in cattle) (EPA, 2021). Although TROPOMI has global coverage, we further focus on a smaller spatial domain centered in northeast Colorado (bounded by

latitudes 39.5 and 41.1 and longitudes -104.5 and -103.7) highlighted in blue in Figure 3.1, as this region contains both of these anthropogenic methane sources. Specifically, it contains a large portion of the Denver-Julesburg (DJ) Basin, a significant oil and gas producing region spanning northeast Colorado, southeast Wyoming, and southwest Nebraska. Additionally, it contains part of Weld County, a major agricultural center with many cattle farms.

The TROPOMI methane data used in this chapter are available from the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) (NASA, 2021).

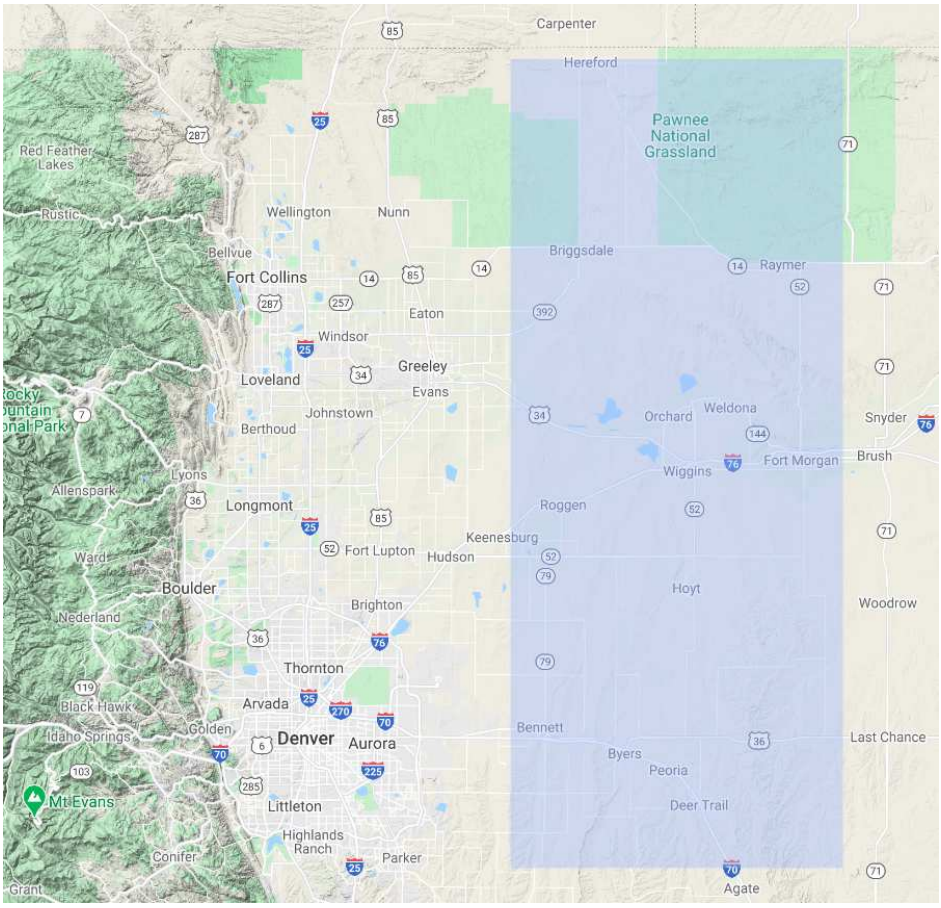

Figure 3.1 Map of the Denver metropolitan area, Boulder, and Fort Collins with our selected region of study highlighted in blue. This region contains oil and gas production and cattle farms, the two largest anthropogenic sources of methane emissions in the United States.

3.3 Hierarchical Spatial Model for Estimating Methane Field

Methane concentrations are assumed to be a continuous field across space. We are interested in estimating this field at specific points (referred to as the "prediction grid") using the TROPOMI methane observations. This type of data analysis can be broadly categorized as a change of support problem. That

is, we are interested in using observations at a coarse spatial resolution (TROPOMI observations, z) to estimate the same quantity at a finer resolution (the continuous methane field sampled at the prediction grid, c). We can write this problem as $z = \mathcal{F}(c)$, where $\mathcal{F}$ is the function that transforms the underlying methane field to the TROPOMI observations. Ultimately, however, we are interested in computing the methane field as a function of the observed data, or $c = \mathcal{F}^{-1}(z)$.

To solve this problem, we have created a hierarchical spatial model. The TROPOMI observations, z, make up the highest level of our model. These observations are a function of the underlying methane field, c, which makes up the second level. To simplify this problem we assume that c is the methane field on a fine and regular grid. The resolution of this grid is set so that the interpolation between grid points does not incur appreciable error. We further assume that the observations, z, are an average of the methane field, c, over the spatial domain of the footprint. Future work will utilize retrieval characteristics to closer approximate the true averaging process, which is more sophisticated than a simple average. We model c as a fixed term plus a spatial process, both of which are governed by a set of parameters that, along with a parameter controlling the measurement error in z, make up the lowest level of our model.

Accordingly, the highest level of our model (the data level) is written as

$$z = Wc + \epsilon, \tag{3.1}$$

where W is a matrix that averages the methane field, c, to produce each entry in z and $\epsilon \sim \mathrm{N}(0, \tau^2)$ is a Gaussian white noise error term. Each row of W applies a weight to every entry in c, such that Wc produces the values of the TROPOMI observations. In this study, we apply equal weight to c within each observation footprint. The W matrix is considered known and fixed, although improvements to this assumption are discussed in Section 3.7. Note that W is not an invertible matrix, which makes directly computing c impossible.

The second level of our model (the process level) is written as

$$c = X\beta + y, \tag{3.2}$$

where X is a matrix of covariates, β is the corresponding vector of coefficients, $y \sim \mathrm{N}(0,\, \sigma^2 K(\theta))$ is a Gaussian process, and $K(\theta)$ is the covariance matrix that results from evaluating the covariance function, $k(s_i, s_j|\theta, \sigma^2)$, at all points in the prediction grid. We assume an exponential covariance function, namely

$$k(s_i, s_j|\theta, \sigma^2) = \sigma^2 \exp\left(-\frac{||s_i - s_j||}{\theta}\right) \tag{3.3}$$

where $||s_i - s_j||$ is the great circle distance between the two points s_i and s_j. Our choice of covariance function is arbitrary at this point, and future work will implement a data-driven method for selecting k. This model definition does not require a specific set of covariates, and we present one modeling option in

Section 3.6. Finally, we assume that $\text{cov}(y, \epsilon) = 0$, or that the spatial process, y, is independent of the measurement error governed by τ^2.

The third level of our model (the parameter level) contains the parameters that govern the data and process levels. We have τ^2, the variance of the measurement error, σ^2, the process variance, θ, the range parameter, and β, the coefficients of the fixed part of the process level model. Instead of a Bayesian approach in which we would specify prior distributions for each of these parameters, we instead estimate them directly via maximum likelihood.

Bringing together the data and process levels and the assumptions discussed above, we get that

$$z = WX\beta + Wy + \epsilon, \tag{3.4}$$

with y and ϵ multivariate normal, or equivalently

$$z \sim \text{N}(WX\beta,\ \sigma^2 WK(\theta)W^T + \tau^2 I). \tag{3.5}$$

3.4 Estimating Model Parameters

We use maximum likelihood to estimate the covariance parameters $\omega = (\sigma^2,\ \tau^2,\ \theta)$ and the coefficients β. We begin by using Equation 3.5 to write the PDF for z given ω and β, and hence the associated log likelihood,

$$\mathcal{L}(\omega, \beta|z) = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma(\omega)|) - \frac{1}{2}(z - A\beta)^T\Sigma(\omega)^{-1}(z - A\beta), \tag{3.6}$$

where m is the number of TROPOMI observations, $A = WX$, and $\Sigma(\omega) = \sigma^2 WK(\theta)W^T + \tau^2 I$ is the covariance of z. We wish to maximize $\mathcal{L}$ over ω and β. We first hold ω fixed and maximize over just β. This results in the generalized least squares (GLS) estimate for β, given by

$$\hat{\beta}(\omega) = (A^T\Sigma(\omega)^{-1}A)^{-1}A^T\Sigma(\omega)^{-1}z. \tag{3.7}$$

We now plug $\hat{\beta}(\omega)$ back into the log likelihood to get

$$\mathcal{L}(\omega, \hat{\beta}(\omega)|z) = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma(\omega)|) - \frac{1}{2}(z - A\hat{\beta}(\omega))^T\Sigma(\omega)^{-1}(z - A\hat{\beta}(\omega)), \tag{3.8}$$

which is solely a function of ω and is called a profiled likelihood. Next we analytically maximize $\mathcal{L}$ over the process variance, σ^2. To do this, we must first reparameterize using a smoothing parameter, defined as $\lambda = \tau^2/\sigma^2$. Under this reparameterization, the covariance matrix becomes

$$\Sigma(\omega) = \sigma^2(WK(\theta)W^T + \lambda I), \tag{3.9}$$

and plugging this expression into Equation 3.7 gives

$$\hat{\beta}(\theta, \lambda) = (A^T(WK(\theta)W^T + \lambda I)^{-1}A)^{-1}A^T(WK(\theta)W^T + \lambda I)^{-1}z, \tag{3.10}$$

which notably does not depend on σ^2. With this profiled estimate for β, we can simplify the third term of the likelihood to make the maximization over σ^2 easier. This term is a weighted residual sum of squares between the observations z and the fixed component of the model $A\beta$. Therefore, we define

$$\text{RSS}(\theta, \lambda) = (z - A\hat{\beta}(\theta, \lambda))^T (WK(\theta)W^T + \lambda I)(z - A\hat{\beta}(\theta, \lambda)), \tag{3.11}$$

which again does not depend on σ^2. Writing the likelihood in terms of $\text{RSS}(\theta, \lambda)$ gives

$$\mathcal{L}(\omega, \hat{\beta}(\omega)|z) = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma(\omega)|) - \frac{\text{RSS}(\theta, \lambda)}{2\sigma^2}. \tag{3.12}$$

Finally, note that

$$|\Sigma(\omega)| = (\sigma^2)^m |WK(\theta)W^T + \lambda I| \tag{3.13}$$

$$\ln(|\Sigma(\omega)|) = m\ln(\sigma^2) + \ln|WK(\theta)W^T + \lambda I|, \tag{3.14}$$

and hence the log likelihood simplifies to

$$\mathcal{L}(\omega, \hat{\beta}(\omega)|z) = -\frac{m}{2}\ln(2\pi) - \frac{m}{2}\ln(\sigma^2) - \frac{1}{2}\ln|WK(\theta)W^T + \lambda I| - \frac{\text{RSS}(\theta, \lambda)}{2\sigma^2}. \tag{3.15}$$

We are now well positioned to maximize σ^2 analytically. Setting the partial of Equation 3.15 equal to zero and solving for σ^2 gives

$$\hat{\sigma}^2(\theta, \lambda) = \frac{\text{RSS}(\theta, \lambda)}{m}. \tag{3.16}$$

We substitute $\hat{\sigma}^2$ back into the log likelihood to get a much reduced expression:

$$\mathcal{L}(\theta, \lambda|z) = -\frac{m}{2}\ln(2\pi) - \frac{m}{2}\ln(\hat{\sigma}^2(\theta, \lambda)) - \frac{1}{2}\ln|WK(\theta)W^T + \lambda I| - \frac{m}{2}. \tag{3.17}$$

The likelihood is now solely a function of the range parameter, θ, and the smoothness parameter, λ, rather than the three covariance parameters, $\omega = (\sigma^2, \tau^2, \theta)$ and the coefficients β. We maximize Equation 3.17 via a simple grid search over the remaining two parameters: θ and λ. For each parameter combination in the grid search, we perform three steps:

- Compute $\hat{\beta}(\theta, \lambda)$ using Equation 3.10 for a sequence of q (typically around 50 or 60) sequential TROPOMI overpasses. Define $\hat{\beta}_{avg}(\theta, \lambda)$ as the average of these q GLS estimates. $\hat{\beta}_{avg}(\theta, \lambda)$ is a more robust estimate than the GLS estimate from a single TROPOMI overpass. Note that we discard TROPOMI overpasses with less than 22 observations (or about 10% coverage of our area of interest), as these overpasses do not have enough data to properly estimate the mean field. Future work might implement a more data-driven approach to this data cleaning step.

- Compute a similarly robust estimate of the process variance, denoted $\hat{\sigma}^2_{avg}(\theta, \lambda)$, using Equation 3.16 and $\hat{\beta}_{avg}(\theta, \lambda)$.

- Compute $\mathcal{L}(\theta, \lambda|z)$ using Equation 3.17 and $\hat{\sigma}^2_{avg}(\theta, \lambda)$ for each of the q sequential overpasses, denoted $\mathcal{L}_i$, where $i = 1, 2, ..., q$. Define $\mathcal{L}_{total} = \sum_{i=1}^{q} \mathcal{L}_i$, which is used to select the optimal values of θ and λ. Future work will explore a method of weighting each $\mathcal{L}_i$ based on the quality of available data for the i^{th} overpass.

We select the values of θ and λ that maximize $\mathcal{L}_{total}$ as the maximum likelihood estimates (MLEs), denoted $\hat{\theta}_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$. We then compute the MLEs for β and σ^2 by evaluating Equations 3.10 and 3.16, respectively, using $\hat{\theta}_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$. Because of the invariance of MLEs under 1-1 transformations, we can easily retrieve the MLE for τ^2 given $\hat{\sigma}^2_{\text{MLE}}$ and $\hat{\lambda}_{\text{MLE}}$ because $\tau^2 = \sigma^2\ \lambda$. With these estimates, we have fully defined our model.

Note that using multiple TROPOMI overpasses to compute $\hat{\beta}_{avg}$ assumes that each overpass is independent of the others. This is a reasonable assumption, but the model could be improved. The seasonal trend in atmospheric methane and weather conditions can introduce dependence in methane concentrations over time. Furthermore, large methane emissions do not always dissipate after 24 hours, meaning that the TROPOMI overpasses following such an emission would be correlated to the day of the emission event. Future work will exploit the correlation between subsequent TROPOMI overpasses, discussed further in Section 3.7.

3.5 Predictions and Uncertainty

With maximum likelihood estimates for the covariance parameters $\omega = (\sigma^2, \tau^2, \theta)$ and the coefficients β, we can now create predictions for the spatial process, y. Using properties from the conditional normal distribution we get that

$$\hat{y} = \text{cov}(y, z)\ \text{cov}(z, z)^{-1}\ (z - WX\hat{\beta}_{\text{MLE}}), \tag{3.18}$$

where the covariance functions here are the same as in Equation 3.3. It is straightforward to then estimate the methane field, which is given by

$$\hat{c} = X\hat{\beta}_{\text{MLE}} + \hat{y}. \tag{3.19}$$

However, we are interested not only in predicting the methane field, but also in quantifying the uncertainty in our prediction. To do this, we use conditional simulation to create an ensemble of equally likely methane fields given the TROPOMI observations, z. We get an estimate of the uncertainty in our predictions by examining the variability within this ensemble. Note that we could have instead computed standard errors for our predictions using properties of the conditional normal given the relatively small size of our study region (see Figure 3.1).

A general outline of the conditional simulation algorithm is given below:

- Using the MLEs $\hat{\theta}_{MLE}$ and $\hat{\lambda}_{MLE}$, we generate a synthetic spatial process, y^*, by computing

$$y^* = L^T u, \tag{3.20}$$

where L is the Cholesky decomposition of $\text{cov}(s_{\text{pred}})$, s_{pred} are the locations on which we predict the methane field, and $u \sim \text{N}(0, 1)$. This result follows from the definition of the Cholesky decomposition and the affine transformation properties of the multivariate normal distribution.

- For each day, we generate synthetic TROPOMI observations, z^*, according to

$$z^* = Wc^* + \epsilon^*, \tag{3.21}$$

where $c^* = X\hat{\beta}_{MLE} + y^*$, ϵ^* is a draw from $\text{N}(0, \hat{\tau}^2_{MLE})$, and W captures the footprint geometry for each overpass.

- Using z^* and the equations discussed earlier in this section, we then compute predictions, $\hat{y}^*$, of the synthetic field, y^*.

- Since we created y^*, we know it exactly. This allows us to compute the error of our prediction, $e = \hat{y}^* - y^*$.

- Finally, we can repeat this process for M many synthetic fields. Each e is an equally likely draw from $y - \hat{y}$, making each $\hat{y} + e$ an equally likely spatial process. The collection of M equally likely processes is called an ensemble.

With this ensemble, we can compute the standard error of the M fields at each prediction location. This gives an estimate of the prediction standard error with variance proportional to $1/M$. We can also use this ensemble to perform any number of interesting inferences, discussed further in the following section.

3.6 Application to TROPOMI Overpass

In this section we apply our modeling framework to a TROPOMI overpass occurring on July 9, 2020. We predict the methane field, c, for this overpass on an equally spaced grid, with predictions occurring every 0.0175 degrees in both latitude and longitude (or about two km in both directions). This grid size was selected to balance prediction fidelity and computational expense. A prediction grid at this resolution results in about 16 prediction locations per TROPOMI footprint for this overpass. Figure 3.2(a) shows our study region for scale and reference. Figure 3.2(b) shows the TROPOMI methane observations within our selected region from the July 9 overpass. The prediction grid we have selected is plotted over the TROPOMI observations.

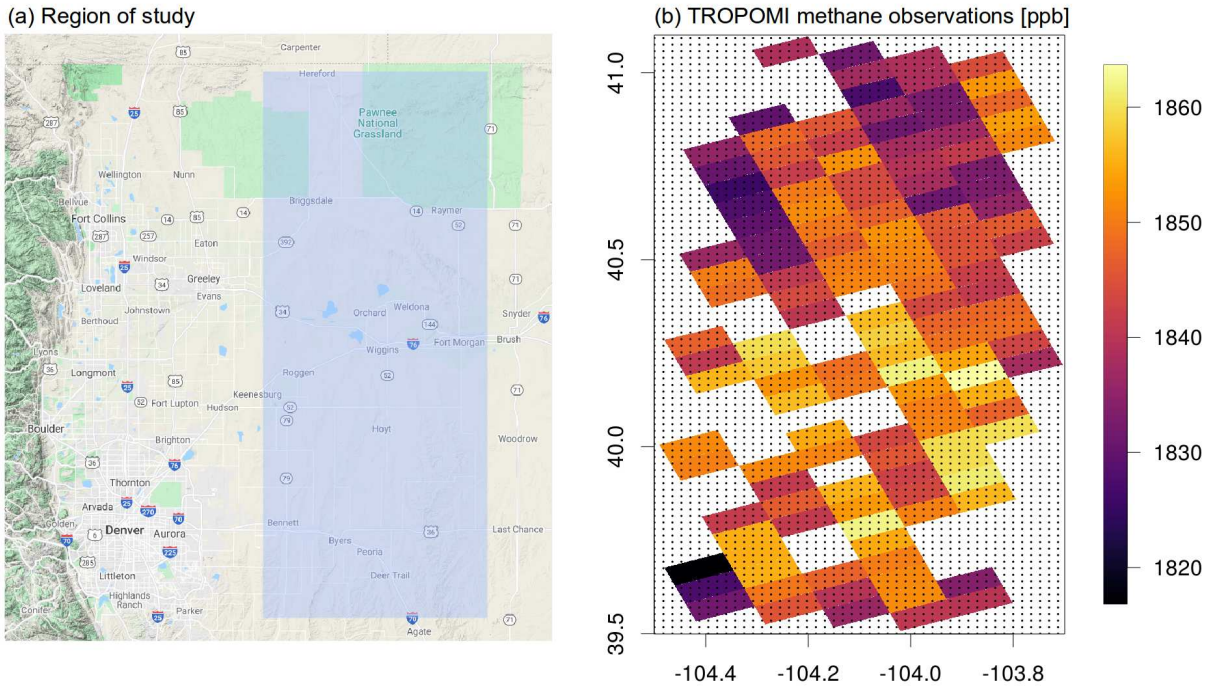

Figure 3.2 (a) Map of the Denver metropolitan area with our selected region of study highlighted in blue. (b) TROPOMI methane observations over our region of study on July 9, 2020 with prediction grid overlaid.

For this example, we have selected four covariates and a fixed term to model the mean trend. Two of the covariates are simply latitude and longitude, which will account for any linear gradients in the field. However, in such a small area, we do not expect these terms to have a large influence on the mean field. The other two covariates account for the two primary anthropogenic sources of methane in the United States: natural gas production and enteric fermentation (i.e. the digestive system in cattle). We expect these terms to have a relatively constant affect on methane concentrations on the scale of two to three months, which is the time frame we consider when estimating model parameters.

We include natural gas production in our model using a relatively simply metric: the number of producing wells within a two km radius (using the same grid as the prediction grid described above). This metric highlights areas with a large oil and gas presence. Well data for Colorado was obtained from the Colorado Oil and Gas Conservation Commission (COGCC), which has a wealth of data related to the oil and gas industry in Colorado (COGCC, 2021). One quantity they record is simply the location of each oil and gas well in the state and its current production status (i.e. is it producing or not producing). We could have also included a covariate related to the volume of oil and gas being produced, rather than simply the presence of wells. However, exploratory analysis showed that these data were extremely heavy tailed, resulting in a small number of prediction grid locations having a mean field estimate orders of magnitude

higher than the rest. Instead of transforming this variable, we simply discarded it, as it highlighted largely the same area as the variable capturing the presence of wells.

We include the influence of enteric fermentation in our model with a similar metric: the number of cattle within a seven km radius (again using the same grid size as the prediction grid). To create this covariate, we use the Gridded Livestock of the World (GLW) data set from the Food and Agriculture Organization of the United Nations. Specifically, we use their GLW3 product, which was created in 2016 and has a much finer resolution than previous products at 0.083 decimals degrees (approximately 10 km at the equator) (Gilbert *et al.*, 2018). This data set is based on a collection of agriculture censuses performed at a variety of spatial scales. These include global censuses performed by the UN, country wide censuses, and occasionally county or province level censuses. In the United States, county level data is available from the USDA Agriculture Census. The data is then down sampled to a much finer resolution using a number of spatial predictors (including anthropogenic, topography, vegetation, and climate variables, often from MODIS) (Gilbert *et al.*, 2018).

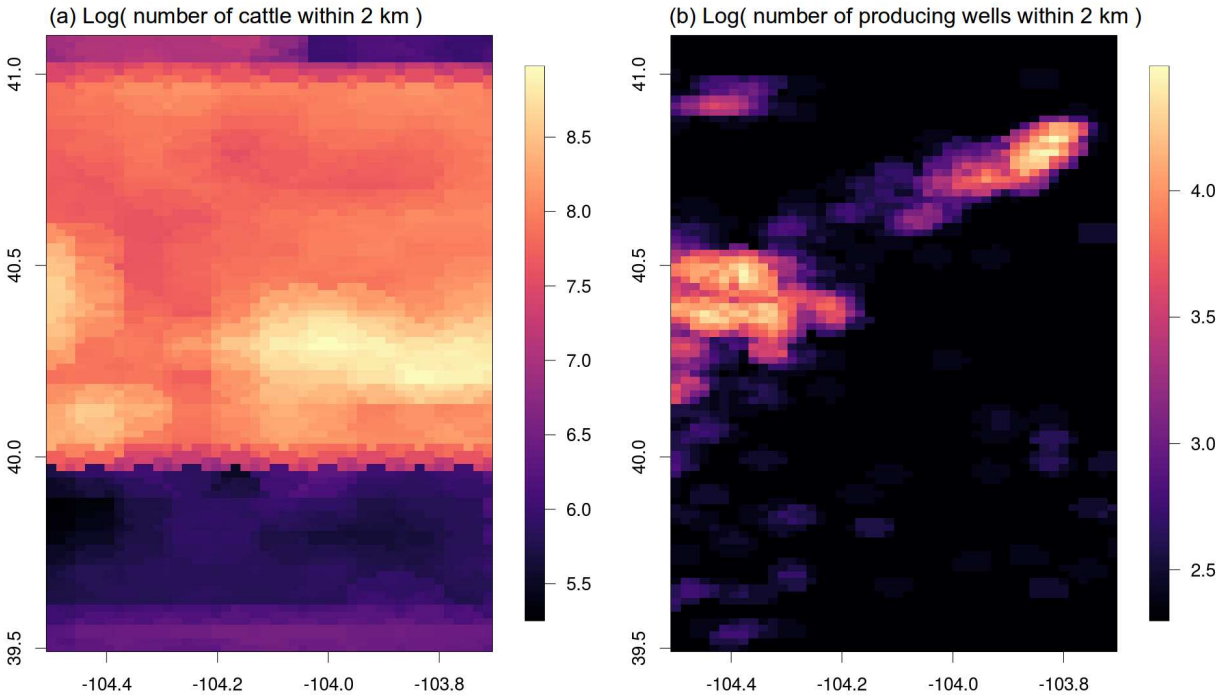

Figure 3.3 The two covariates related to anthropogenic sources of methane plotted over space. (a) Log of the number of cattle within two km of each prediction grid point. Data from the GLW3 product (Gilbert *et al.*, 2018). (b) Log of the number of producing oil and gas wells within two km of each prediction grid point. Data from the COGCC (COGCC, 2021).

Figure 3.3 shows the two covariates related to anthropogenic sources of methane described above. Note that we take the log of the count data, as the counts were heavy tailed. Exploratory analysis found that

log transforming these variables results in smoother mean field estimates. Figure 3.3(a) shows the log of the number of cattle within two km of each prediction grid point. Note that there is a fairly sharp boundary at 40 degrees of latitude. This is a result of the down sampling used in the GLW3 product. This line of latitude was likely a boundary of one of the MODIS pixels, resulting in a sharp difference in cattle estimates on either side. Future work will consider different ways of modeling the cattle data. Figure 3.3(b) shows the log of the number of producing wells within two km of each prediction grid point. Note that a constant value of 10 was applied to the count data before taking the log to account for zero counts.

With these covariate definitions, we can estimate model parameters as described in Section 3.4. For this example, we use 67 TROPOMI overpasses occurring between May and August of 2020 to estimate the MLEs. The log likelihood surface resulting from the grid search over λ and θ is shown in Figure 3.4. Here we plot the $\mathcal{L}_{total}$ values resulting from each combination of λ and θ values.

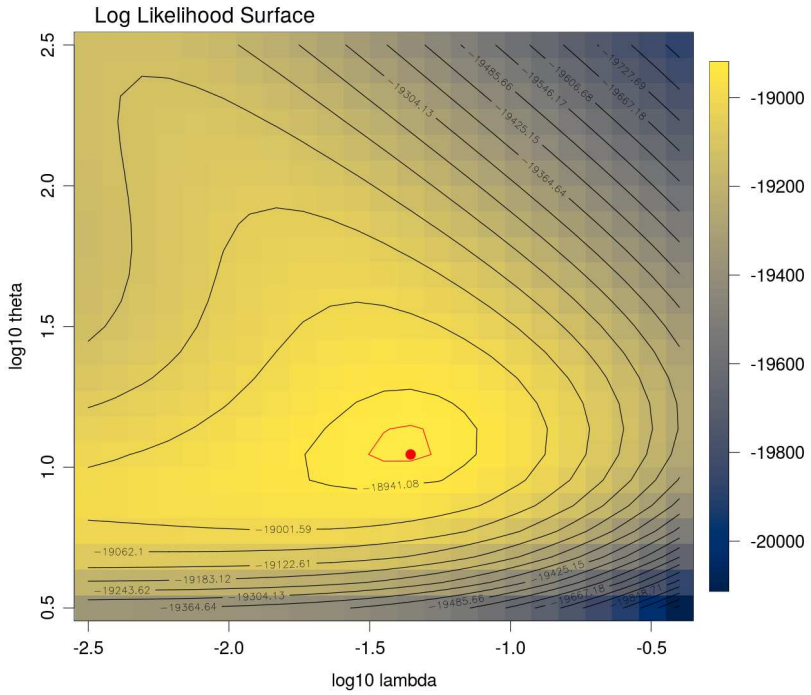

Figure 3.4 Log likelihood surface used to estimate λ and θ. The MLEs are shown as a red dot and a 95% confidence level is drawn in red.

Figure 3.5(b) shows the estimated mean field after fitting β as described in Section 3.4. The presence of oil and gas wells is clearly reflected in the mean field estimate, but the presence of cattle is largely ignored. This is because the size of the estimated coefficient on oil and gas wells was much larger than the estimated coefficient on cattle, which was nearly zero. This could be because the cattle data was too coarse to provide any useful information. There is also a linear gradient from the latitude and longitude terms.

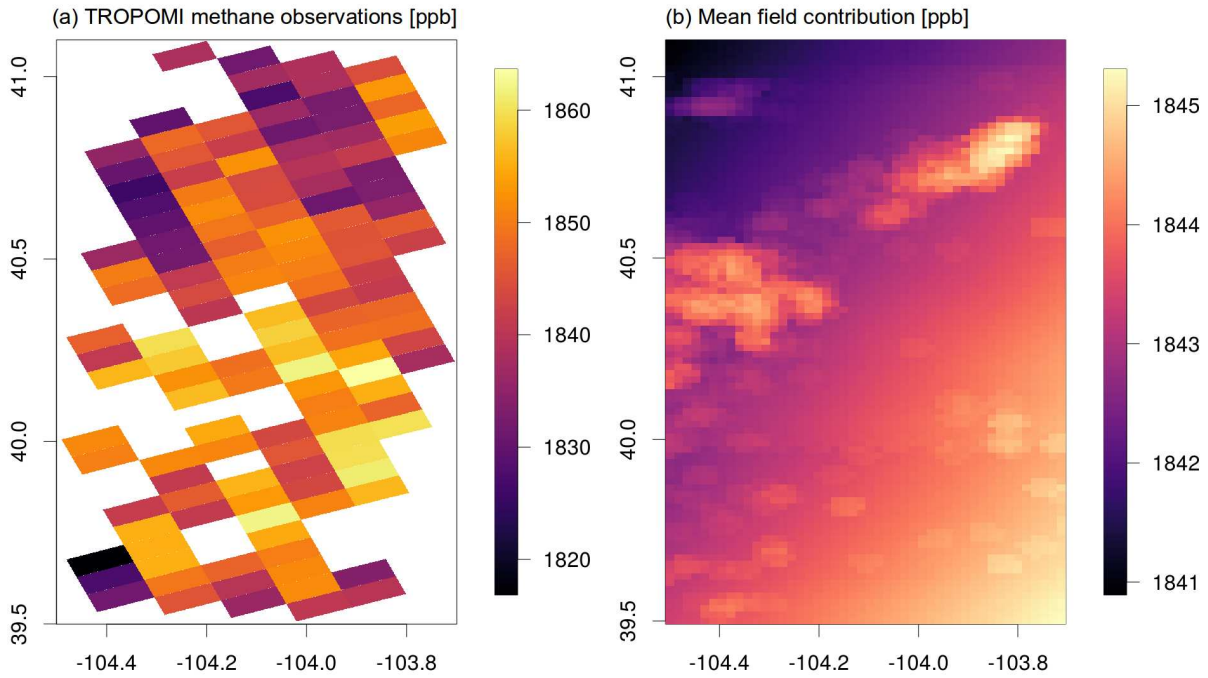

Figure 3.5 (a) TROPOMI observations on July 9, 2020 for reference. (b) Mean field contribution based on $\hat{\beta}_{MLE}$. Note the dramatically different scales between (a) and (b), which indicates that the mean field is not important in the overall methane field estimate for such a small spatial domain.

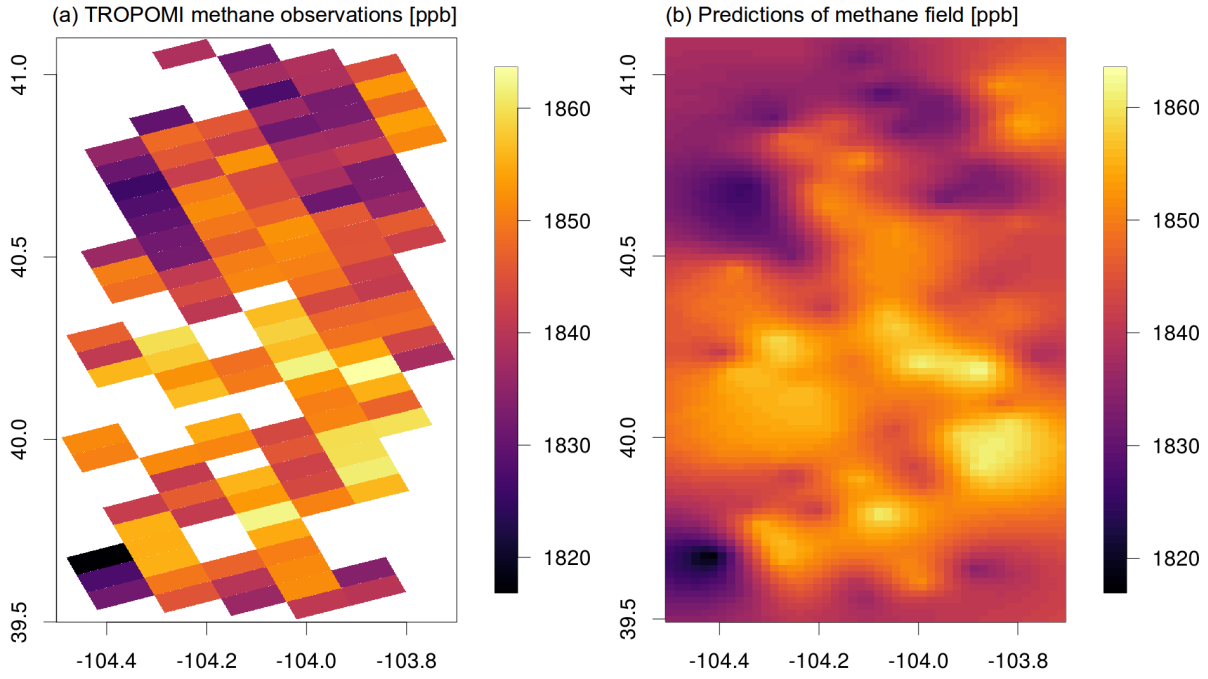

Figure 3.6 (a) TROPOMI methane observations in our study region on July 9, 2020. (b) Estimated methane field given the TROPOMI observations and MLEs.

It is important to note that the methane scale in Figure 3.5(b) is much smaller than that of Figure 3.5(a). This indicates that, ultimately, the mean field is not important in the methane field estimation. This makes sense, as we are working with a very small region, so we do not expect there to be a very noticeable mean trend to remove. Note that in Figure 3.6 (discussed momentarily), the range of the estimated methane field is much larger than the range of the mean field shown in Figure 3.5(b).

With maximum likelihood estimates for the covariance parameters $\omega = (\sigma^2, \tau^2, \theta)$ and the coefficients β, we can use Equations 3.18 and 3.19 to predict the continuous methane concentrations on the prediction grid described above. Figure 3.6(a) again shows the TROPOMI observations and their respective footprints for the overpass on July 9, 2020. Figure 3.6(b) shows predictions of the methane field.

Finally, we perform the conditional simulation with an ensemble size of $M = 200$. Prediction standard errors are shown in Figure 3.7(a). The blue points show the center of each TROPOMI footprint. The standard errors clearly increase as you move away from the observations. Figure 3.7(a) highlights an overpass with relatively dense observations. More frequently, however, TROPOMI observations are quite sparse (usually due to cloud cover). These standard error plots will be especially useful in scenarios with less data.

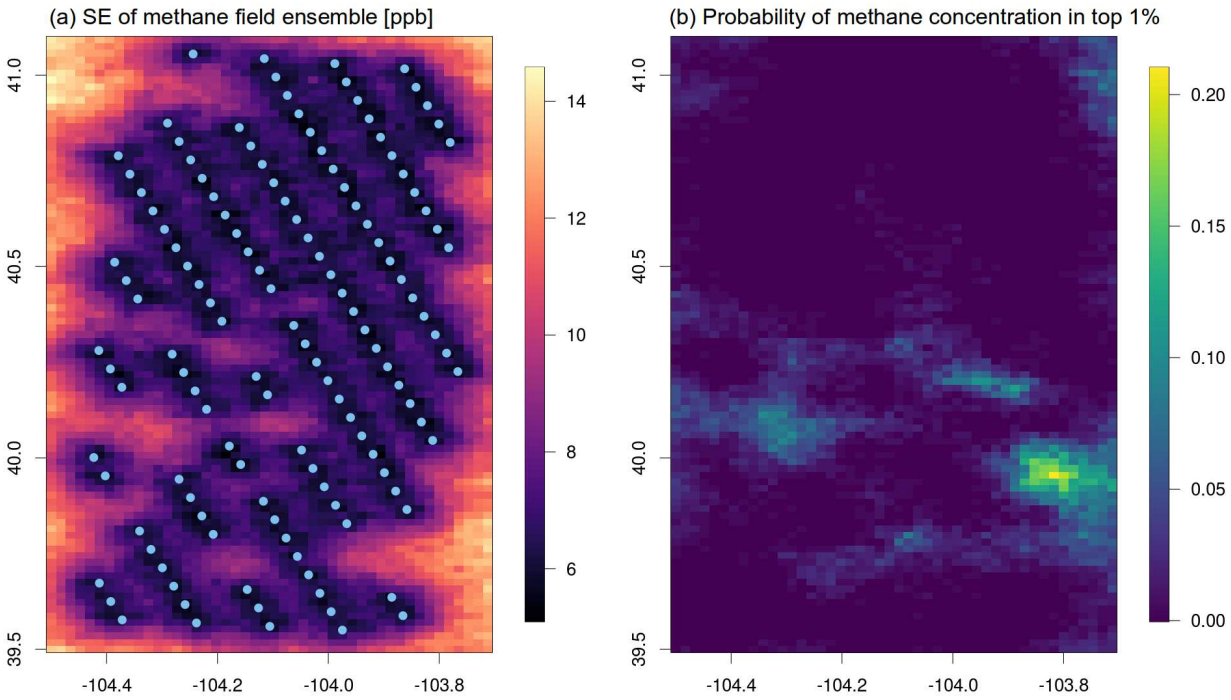

Figure 3.7 (a) Standard errors from an ensemble with $M = 200$. (b) Estimated probability of each prediction location containing a methane concentration in the top 1% of all predictions within the ensemble.

In addition to giving us an estimate of prediction uncertainty, the ensemble of equally likely methane fields allows us to perform other interesting inferences. Figure 3.7(b) highlights one example. Here we plot the estimated probability of a prediction location containing a methane concentration in the top 1% of all conditional means within the ensemble. These probabilities clearly highlight three or four locations that are most likely to contain the highest methane concentrations in the region for this given overpass. This type of inference is of particular interest, as a large percent of anthropogenic methane emissions comes from a small percent of emission events (Zavala-Araiza *et al.*, 2017). It is therefore important to locate these "super emitter events" so that their sources can be identified and addressed, hence reducing future emissions.

3.7 Summary

We are interested in predicting continuous methane concentrations on a fine grid given remotely sensed observations from the TROPOMI instrument. Methane is a potent greenhouse gas, and a better understanding of the continuous methane field across an oil and gas producing basin can help identify emission events. We have created a hierarchical spatial model for estimating the underlying methane field from a set of TROPOMI observations, largely following the model definition from Nguyen *et al.* (2017). This framework takes into account the footprint of the TROPOMI observations when estimating model parameters, rather than treating them as point concentrations. We use maximum likelihood to estimate model parameters and account for multiple TROPOMI overpasses to ensure that these estimates are robust. Furthermore, we use conditional simulation to create an ensemble of equally likely methane fields given a single TROPOMI overpass. This ensemble allows us to quantify the uncertainty in our estimates and perform interesting inferences. As an example, we use the ensemble to compute the probability of each location within the study region containing a methane concentration in the top 1% of all predictions within the ensemble. This highlights a small number of locations that are most likely to contain the largest methane concentrations.

There are many ways in which we could potentially improve the predictions from this model, with some listed below.

- Use cross-validation to select the smoothness of the covariance function, as opposed to arbitrarily selecting an exponential form. A smoother member of the Matérn family might be better suited to these data.
- Incorporate temporal dependence in the q subsequent methane fields used to estimate $\hat{\beta}$.
- "Sharpen" the W matrix by including fractional weights when a prediction location is partly inside of a given TROPOMI footprint.

- Add weights to the W matrix to account for specific aspects of the TROPOMI retrieval algorithm. This would better approximate the true "forward model" that converts the methane concentrations to TROPOMI observations. Other options would be to weight the measurement error or include a bias term.
- Implement a fully Bayesian approach in which we specify prior distributions for each parameter and sample the resulting posterior with Markov chain Monte Carlo (MCMC) methods.
- Model the influence of advection and diffusion by adding vector wind fields to the model.

Future work will also investigate assimilation techniques for constraining the output of this model with data from the other emissions monitoring techniques described in Section 3.1. In particular, we have access to continuously monitored methane data from Project Canary, an environmental standards company. These data are collected in near real time on oil and gas sites across Colorado.

Ultimately, we believe that the modeling framework presented here is a first step towards extracting meaningful information on a small scale using only the coarsely pixelated TROPOMI observations. We hope that this work will help monitor methane emissions via satellite and provide avenues for improvement and comparison to more localized, ground-based sensors.

CHAPTER 4

CONCLUSION

We have presented two modeling studies aimed at addressing pressing environmental issues through the use of remotely sensed data. The first seeks to explain the relationship between climate variability and fire season intensity. We do this by modeling remotely sensed carbon monoxide, a proxy for fire intensity in the Southern Hemisphere. These models are parsimonious by design, allowing for scientific interpretation of the selected climate indices and lag values. We identify the Nino 3.4 climate index lagged at four weeks as a primary driver of atmospheric CO in the Maritime Southeast Asia region. Other important climate indices are the DMI and OLR (as a proxy for the MJO). We further identify that Nino 3.4 interactions with the OLR and DMI are significant predictors, suggesting that the effect of these indices are amplified when they are in phase.

We develop a framework for assessing the stability of the selected model terms, ultimately finding that the terms present in the most parsimonious model are very likely to remain in models refit to resampled training data. This justifies assigning scientific weight to the selection of these model terms, as it suggests that their inclusion in the model is not an artifact of the specific training data used. We show that the models for Maritime Southeast Asia are able to explain about 70% of the variability in weekly CO anomalies. Finally, we show that our models are still able to explain about 65% of the variability in CO when forced to use lags of 35 weeks or greater. This is promising, as it indicates that predictions made relatively far in advance can still capture the overall structure and amplitude of the CO anomalies.

The second study predicts methane concentration on a fine grid using the spatially averaged TROPOMI observations. Because the prediction grid is at a finer resolution than the TROPOMI footprints, we take into account the shape of each footprint when estimating model parameters. We fit a hierarchical spatial model via maximum likelihood and use conditional simulation for uncertainty quantification and inferences on derived quantities. As an example, we estimate the probability of a prediction location containing a methane concentration in the top 1% of all predictions within the ensemble. These probabilities highlight three or four locations likely to contain the highest methane concentrations in the region for this given overpass.

Ultimately, we believe that this modeling framework is a first step towards extracting meaningful information on a small scale using only the TROPOMI observations. We plan on comparing the predictions from our model to other monitoring techniques, such as the site-level continuous monitoring data from Project Canary or the aircraft-level methane data from a JPL overflight campaign planned for

the summer of 2021. We hope that this work will help monitor methane emissions via satellite and provide avenues for improvement and comparison to more localized sensors.

To conclude, we briefly connect the two projects through their mutual dependence on remotely sensed information. For the fire season intensity study, both predictor and response variables in our model depend on remotely sensed data. The response (carbon monoxide) is observed via satellite, and many of the predictors (climate mode indices) are computed using sea surfaces temperature also observed via satellite. For the methane field prediction study, both predictor and response variables again depend on remotely sensed data. The response (methane) is observed via satellite, and one of the predictors (down-sampled cattle counts) depends on remotely sensed vegetation and ground cover information. Clearly, satellite remote sensing provides vital information for environmental studies, as we demonstrate with these two examples.

REFERENCES

Alencar, Ane, Asner, Gregory P., Knapp, David, & Zarin, Daniel. 2011. Temporal variability of forest fires in eastern Amazonia. *Ecological Applications*, **21**(7), 2397–2412.

Andela, Niels, & Van Der Werf, Guido R. 2014. Recent trends in African fires driven by cropland expansion and El Niño to la Niña transition. *Nature Climate Change*, **4**(9), 791–795.

Andreoli, Rita Valéria, & Kayano, Mary Toshie. 2006. Tropical Pacific and South Atlantic effects on rainfall variability over Northeast Brazil. *International Journal of Climatology*, **26**(13), 1895–1912.

Bamston, Anthony G., Chelliah, Muthuvel, & Goldenberg, Stanley B. 1997. Documentation of a highly enso-related sst region in the equatorial pacific: Research note. *Atmosphere - Ocean*, **35**(3), 367–383.

Bien, Jacob, Taylor, Jonathan, & Tibshirani, Robert. 2013. A lasso for hierarchical interactions. *The Annals of Statistics*, **41**(3), 1111–1141.

Breheny, Patrick, & Huang, Jian. 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, **5**(1), 232–253.

Buchholz, R. R., Hammerling, D., Worden, H. M., Deeter, M. N., Emmons, L. K., Edwards, D. P., & Monks, S. A. 2018. Links Between Carbon Monoxide and Climate Indices for the Southern Hemisphere and Tropical Fire Regions. *Journal of Geophysical Research: Atmospheres*, **123**(17), 9786–9800.

Buchwitz, Michael, Schneising, Oliver, Reuter, Maximilian, Heymann, Jens, Krautwurst, Sven, Bovensmann, Heinrich, Burrows, John P., Boesch, Hartmut, Parker, Robert J., Somkuti, Peter, Detmers, Rob G., Hasekamp, Otto P., Aben, Ilse, Butz, André, Frankenberg, Christian, & Turner, Alexander J. 2017. Satellite-derived methane hotspot emission estimates using a fast data-driven method. *Atmospheric Chemistry and Physics*, **17**(9), 5751–5774.

Ceccato, P., Nengah Surati Jaya, I., Qian, J. H., Tippett, M. K., Robertson, A. W., & Someshwar, S. 2010. *Early Warning and Response to Fires in Kalimantan, Indonesia*. Tech. rept. International Research Institute for Climate and Society.

Chen, Yang, Morton, Douglas C., Andela, Niels, Giglio, Louis, & Randerson, James T. 2016. How much global burned area can be forecast on seasonal time scales using sea surface temperatures? *Environmental Research Letters*, **11**(4), 45001.

Cleverly, James, Eamus, Derek, Luo, Qunying, Coupe, Natalia Restrepo, Kljun, Natascha, Ma, Xuanlong, Ewenz, Cacilia, Li, Longhui, Yu, Qiang, & Huete, Alfredo. 2016. The importance of interacting climate modes on Australia's contribution to global carbon cycle extremes. *Scientific Reports*, **6**(1), 1–10.

COGCC. 2021. *Colorado Oil and Gas Conservation Commission*. https://cogcc.state.co.us/#/home. (Accessed on 04/21/2021).

Cressie, Noel. 1993. *Statistics for Spatial Data*. Wiley.

Cressie, Noel, & Johannesson, Gardar. 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 209–226.

Crowell, Sean, Baker, David, Schuh, Andrew, Basu, Sourish, Jacobson, Andrew R., Chevallier, Frederic, Liu, Junjie, Deng, Feng, Feng, Liang, McKain, Kathryn, Chatterjee, Abhishek, Miller, John B., Stephens, Britton B., Eldering, Annmarie, Crisp, David, Schimel, David, Nassar, Ray, O'Dell, Christopher W., Oda, Tomohiro, Sweeney, Colm, Palmer, Paul I., & Jones, Dylan B.A. 2019. The 2015-2016 carbon cycle as seen from OCO-2 and the global in situ network. *Atmospheric Chemistry and Physics*, **19**(15), 9797–9831.

Deeter, M. N., Edwards, D. P., Gille, J. C., & Drummond, James R. 2007. Sensitivity of MOPITT observations to carbon monoxide in the lower troposphere. *Journal of Geophysical Research*, **112**(D24), D24306.

Deeter, M. N., Martínez-Alonso, S., Edwards, D. P., Emmons, L. K., Gille, J. C., Worden, H. M., Sweeney, C., Pittman, J. V., Daube, B. C., & Wofsy, S. C. 2014. The MOPITT Version 6 product: Algorithm enhancements and validation. *Atmospheric Measurement Techniques*, **7**(11), 3623–3632.

Deeter, Merritt N., Edwards, David P., Francis, Gene L., Gille, John C., Mao, Debbie, Martínez-Alonso, Sara, Worden, Helen M., Ziskin, Dan, Andreae, Meinrat O., & Andreae, Meinrat O. 2019. Radiance-based retrieval bias mitigation for the MOPITT instrument: The version 8 product. *Atmospheric Measurement Techniques*, **12**(8), 4561–4580.

Edwards, D. P., Emmons, L. K., Gille, J. C., Chu, A., Attié, J.-L., Giglio, L., Wood, S. W., Haywood, J., Deeter, M. N., Massie, S. T., Ziskin, D. C., & Drummond, J. R. 2006a. Satellite-observed pollution from Southern Hemisphere biomass burning. *Journal of Geophysical Research*, **111**(D14), D14312.

Edwards, D. P., Pétron, G., Novelli, P. C., Emmons, L. K., Gille, J. C., & Drummond, J. R. 2006b. Southern Hemisphere carbon monoxide interannual variability observed by Terra/Measurement of Pollution in the Troposphere (MOPITT). *Journal of Geophysical Research*, **111**(D16), D16303.

EIA. 2020 (Jun). *U.S. Energy Information Administration - How much carbon dioxide is produced when different fuels are burned?* https://www.eia.gov/tools/faqs/faq.php?id=73&t=11. (Accessed on 12/16/2020).

Enfield, David B., Mestas-Nuñez, Alberto M., Mayer, Dennis A., & Cid-Serrano, Luis. 1999. How ubiquitous is the dipole relationship in tropical Atlantic sea surface temperatures? *Journal of Geophysical Research: Oceans*, **104**(C4), 7841–7848.

EPA. 2020 (Sep). *U.S. Environmental Protection Agency - Understanding Global Warming Potentials.* https://www.epa.gov/ghgemissions/understanding-global-warming-potentials. (Accessed on 12/16/2020).

EPA. 2021 (April). *Environmental Protection Agency - Greenhouse Gas Emissions — Methane Emissions.* https://www.epa.gov/ghgemissions/overview-greenhouse-gases#methane. (Accessed on 04/20/2021).

Field, Robert D., Van Der Werf, Guido R., Fanin, Thierry, Fetzer, Eric J., Fuller, Ryan, Jethva, Hiren, Levy, Robert, Livesey, Nathaniel J., Luo, Ming, Torres, Omar, & Worden, Helen M. 2016. Indonesian fire activity and smoke pollution in 2015 show persistent nonlinear sensitivity to El Niño-induced drought. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(33), 9204–9209.

Fox, Thomas A., Barchyn, Thomas E., Risk, David, Ravikumar, Arvind P., & Hugenholtz, Chris H. 2019 (may). *A review of close-range and screening technologies for mitigating fugitive methane emissions in upstream oil and gas.*

Fuller, Douglas O., & Murphy, Kevin. 2006. The ENSO-fire dynamic in insular Southeast Asia. *Climatic Change*, **74**(4), 435–455.

Giglio, Louis, Schroeder, Wilfrid, & Justice, Christopher O. 2016. The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sensing of Environment*, **178**(jun), 31–41.

Giglio, Louis, Boschetti, Luigi, Roy, David P., Humber, Michael L., & Justice, Christopher O. 2018. The Collection 6 MODIS burned area mapping algorithm and product. *Remote Sensing of Environment*, **217**(nov), 72–85.

Gilbert, Marius, Nicolas, Gaëlle, Cinardi, Giusepina, Van Boeckel, Thomas P., Vanwambeke, Sophie O., Wint, G. R.William, & Robinson, Timothy P. 2018. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Scientific Data*, **5**(1), 1–11.

Hao, Ning, Feng, Yang, & Zhang, Hao Helen. 2018. Model Selection for High-Dimensional Quadratic Regression via Regularization. *Journal of the American Statistical Association*, **113**(522), 615–625.

Hirst, Bill, Jonathan, Philip, González del Cueto, Fernando, Randell, David, & Kosut, Oliver. 2013. Locating and quantifying gas emission sources using remotely obtained concentration data. *Atmospheric Environment*, **74**(aug), 141–158.

Holloway, Tracey, Levy, Hiram, & Kasibhatla, Prasad. 2000. Global distribution of carbon monoxide. *Journal of Geophysical Research: Atmospheres*, **105**(D10), 12123–12147.

Hu, Haili, Hasekamp, Otto, Butz, André, Galli, André, Landgraf, Jochen, Aan De Brugh, Joost, Borsdorff, Tobias, Scheepmaker, Remco, & Aben, Ilse. 2016. The operational methane retrieval algorithm for TROPOMI. *Atmospheric Measurement Techniques*, **9**(11), 5423–5440.

Jacob, Daniel J., Turner, Alexander J., Maasakkers, Joannes D., Sheng, Jianxiong, Sun, Kang, Liu, Xiong, Chance, Kelly, Aben, Ilse, McKeever, Jason, & Frankenberg, Christian. 2016. Satellite observations of atmospheric methane and their value for quantifying methane emissions. *Atmospheric Chemistry and Physics*, **16**(22), 14371–14396.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, Roy, & Joseph, Dennis. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, **77**(3), 437 – 472.

Kistler, Robert, Kalnay, Eugenia, Collins, William, Saha, Suranjana, White, Glenn, Woollen, John, Chelliah, Muthuvel, Ebisuzaki, Wesley, Kanamitsu, Masao, Kousky, Vernon, van den Dool, Huug, Jenne, Roy, & Fiorino, Michael. 2001. The NCEP–NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bulletin of the American Meteorological Society*, **82**(2), 247–268.

Kort, Eric A., Frankenberg, Christian, Costigan, Keeley R., Lindenmaier, Rodica, Dubey, Manvendra K., & Wunch, Debra. 2014. Four corners: The largest US methane anomaly viewed from space. *Geophysical Research Letters*, **41**(19), 6898–6903.

Levelt, Pieternel F., Van Den Oord, Gijsbertus H.J., Dobber, Marcel R., Mälkki, Anssi, Visser, Huib, De Vries, Johan, Stammes, Piet, Lundell, Jens O.V., & Saari, Heikki. 2006. The ozone monitoring instrument. *IEEE Transactions on Geoscience and Remote Sensing*, **44**(5), 1093–1100.

Madden, Roland A., & Julian, Paul R. 1972. Description of Global-Scale Circulation Cells in the Tropics with a 40–50 Day Period. *Journal of Atmospheric Sciences*, **29**(6), 1109 – 1123.

Madden, Roland A., & Julian, Paul R. 1994. Observations of the 40–50-Day Tropical Oscillation—A Review. *Monthly Weather Review*, **122**(5), 814 – 837.

Munro, Rosemary, Lang, Rudiger, Klaes, Dieter, Poli, Gabriele, Retscher, Christian, Lindstrot, Rasmus, Huckle, Roger, Lacan, Antoine, Grzegorski, Michael, Holdak, Andriy, Kokhanovsky, Alexander, Livschitz, Jakob, & Eisinger, Michael. 2016. The GOME-2 instrument on the Metop series of satellites: Instrument design, calibration, and level 1 data processing - An overview. *Atmospheric Measurement Techniques*, **9**(3), 1279–1301.

NASA. 2021. *Goddard Earth Sciences Data and Information Services Center (GES DISC)*. https://disc.gsfc.nasa.gov/. (Accessed on 05/06/2021).

Neelin, J. David, Battisti, David S., Hirst, Anthony C., Jin, Fei Fei, Wakata, Yoshinobu, Yamagata, Toshio, & Zebiak, Stephen E. 1998. ENSO theory. *Journal of Geophysical Research: Oceans*, **103**(C7), 14261–14290.

Nelder, J. A. 1977. A Reformulation of Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **140**(1), 48.

Nguyen, Hai, Cressie, Noel, & Braverman, Amy. 2017. Multivariate Spatial Data Fusion for Very Large Remote Sensing Datasets. *Remote Sensing*, **9**(2), 142.

NOAA CPC. 2021. *Climate Prediction Center - Teleconnections: Antarctic Oscillation*. https://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/aao/aao.shtml. (Accessed on 04/12/2021).

NOAA OOPC. 2021. *Ocean Observations Panel for Climate - State of the Ocean Climate*. https://stateoftheocean.osmc.noaa.gov/. (Accessed on 04/12/2021).

NOAA PSL. 2021. *Physical Sciences Laboratory - Interpolated OLR*. https://psl.noaa.gov/data/gridded/data.interp_OLR.html. (Accessed on 04/12/2021).

Nur'utami, Murni Ngestu, & Hidayat, Rahmat. 2016. Influences of IOD and ENSO to Indonesian Rainfall Variability: Role of Atmosphere-ocean Interaction in the Indo-pacific Sector. *Procedia Environmental Sciences*, **33**(jan), 196–203.

Nychka, Douglas, Wikle, Christopher, & Royle, J Andrew. 2002. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, **2**(4), 315–331.

Reid, J. S., Xian, P., Hyer, E. J., Flatau, M. K., Ramirez, E. M., Turk, F. J., Sampson, C. R., Zhang, C., Fukada, E. M., & Maloney, E. D. 2012. Multi-scale meteorological conceptual analysis of observed active fire hotspot activity and smoke optical depth in the Maritime Continent. *Atmospheric Chemistry and Physics*, **12**(4), 2117–2147.

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., & Yamagata, T. 1999. A dipole mode in the tropical Indian ocean. *Nature*, **401**(6751), 360–363.

Saji, NH, & Yamagata, T. 2003. Possible impacts of Indian Ocean Dipole mode events on global climate. *Climate Research*, **25**(2), 151–169.

Shabbar, Amir, Skinner, Walter, & Flannigan, Mike D. 2011. Prediction of seasonal forest fire severity in Canada from large-scale climate patterns. *Journal of Applied Meteorology and Climatology*, **50**(4), 785–799.

Shawki, Dilshad, Field, Robert D., Tippett, Michael K., Saharjo, Bambang Hero, Albar, Israr, Atmoko, Dwi, & Voulgarakis, Apostolos. 2017. Long-Lead Prediction of the 2015 Fire and Haze Episode in Indonesia. *Geophysical Research Letters*, **44**(19), 9996.

Thompson, David W. J., & Wallace, John M. 2000. Annular Modes in the Extratropical Circulation. Part I: Month-to-Month Variability. *Journal of Climate*, **13**(5), 1000 – 1016.

Trenberth, KE. 2013. *El Nino Southern Oscillation (ENSO)*. Tech. rept. National Center for Atmospheric Research (NCAR).

Turner, A. J., Jacob, D. J., Wecht, K. J., Maasakkers, J. D., Lundgren, E., Andrews, A. E., Biraud, S. C., Boesch, H., Bowman, K. W., Deutscher, N. M., Dubey, M. K., Griffith, D. W.T., Hase, F., Kuze, A., Notholt, J., Ohyama, H., Parker, R., Payne, V. H., Sussmann, R., Sweeney, C., Velazco, V. A., Warneke, T., Wennberg, P. O., & Wunch, D. 2015. Estimating global and North American methane emissions with high spatial resolution using GOSAT satellite data. *Atmospheric Chemistry and Physics*, **15**(12), 7049–7069.

van der Werf, Guido R., Randerson, James T., Giglio, Louis, Gobron, Nadine, & Dolman, A. J. 2008. Climate controls on the variability of fires in the tropics and subtropics. *Global Biogeochemical Cycles*, **22**(3), n/a–n/a.

Varon, D. J., McKeever, J., Jervis, D., Maasakkers, J. D., Pandey, S., Houweling, S., Aben, I., Scarpelli, T., & Jacob, D. J. 2019. Satellite Discovery of Anomalously Large Methane Point Sources From Oil/Gas Production. *Geophysical Research Letters*, **46**(22), 13507–13516.

Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H. J., de Haan, J. F., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., Vink, R., Visser, H., & Levelt, P. F. 2012. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, **120**(may), 70–83.

Voulgarakis, Apostolos, Marlier, Miriam E., Faluvegi, Greg, Shindell, Drew T., Tsigaridis, Kostas, & Mangeon, Stéphane. 2015. Interannual variability of tropospheric trace gases and aerosols: The role of biomass burning emissions. *Journal of Geophysical Research: Atmospheres*, **120**(14), 7157–7173.

Wheeler, Matthew C., & Hendon, Harry H. 2004. An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. *Monthly Weather Review*, **132**(8), 1917 – 1932.

Wooster, M. J., Perry, G. L.W., & Zoumas, A. 2012. Fire, drought and El Niño relationships on Borneo (Southeast Asia) in the pre-MODIS era (1980-2000). *Biogeosciences*, **9**(1), 317–340.

Wunch, Debra, Toon, Geoffrey C., Blavier, Jean François L., Washenfelder, Rebecca A., Notholt, Justus, Connor, Brian J., Griffith, David W.T., Sherlock, Vanessa, & Wennberg, Paul O. 2011. The total carbon column observing network. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **369**(1943), 2087–2112.

Xavier, Prince, Rahmat, Raizan, Cheong, Wee Kiong, & Wallace, Emily. 2014. Influence of Madden-Julian Oscillation on Southeast Asia rainfall extremes: Observations and predictability. *Geophysical Research Letters*, **41**(12), 4406–4412.

Zavala-Araiza, Daniel, Alvarez, Ramón A., Lyon, David R., Allen, David T., Marchese, Anthony J., Zimmerle, Daniel J., & Hamburg, Steven P. 2017. Super-emitters in natural gas infrastructure are caused by abnormal process conditions. *Nature Communications*, **8**(1), 14012.

Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**(2), 894–942.

Zhang, Yuzhong, Gautam, Ritesh, Pandey, Sudhanshu, Omara, Mark, Maasakkers, Joannes D., Sadavarte, Pankaj, Lyon, David, Nesser, Hannah, Sulprizio, Melissa P., Varon, Daniel J., Zhang, Ruixiong, Houweling, Sander, Zavala-Araiza, Daniel, Alvarez, Ramon A., Lorente, Alba, Hamburg, Steven P., Aben, Ilse, & Jacob, Daniel J. 2020. Quantifying methane emissions from the largest oil-producing basin in the United States from space. *Science Advances*, **6**(17), eaaz5120.