UNIVERSITY OF CALIFORNIA SANTA

CRUZ

DETECTING AND PREDICTING HOT MOMENTS OF METHANE EMISSIONS FROM COASTAL WETLANDS

A thesis submitted in partial satisfaction of
the requirements for the degree of

MASTER OF SCIENCE

in

EARTH SCIENCES

by

Grace D Pearsall

March 2024

The Thesis of Grace Pearsall is
approved:

_____
Professor Adina Paytan, Chair

_____
Professor Claudie Beaulieu

_____
Professor Ariane Arias-Ortiz

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures and Tables

Acknowledgments

Thank you to my advisor, Dr. Adina Paytan, for her guidance, mentorship, understanding, and no-nonsense attitude. Thank you, Adina, for believing in my abilities and letting me take this project wherever I wanted to go. Thanks, Adina, for encouraging me to pursue projects outside of my thesis, such as setting up the eddy covariance towers in Elkhorn Slough and participating in the DIG summer camp. These projects ended up being some of my favorite memories from graduate school. I also want to thank my lovely committee members, Dr. Claudie Beaulieu and Dr. Ariane Arias-Ortiz. Without Claudie and Ari, this project would never have gotten off the ground. I learned so much from both of you, and I am extremely grateful that you shared your skills and were so willing to help with this project. Thank you to everyone at the Biometeorology Lab at UC Berkeley (Daphne Szutu, Joe Verafaillie, and Dr. Camilo Rey-Sanchez) for their help with all things eddy covariance. I'd also like to give a shoutout to the wonderful Paytan Lab group. Thanks to everyone in our lab for their support and some memorable Halloween costumes and white elephant parties. Finally, thank you to my family and friends for all your love and support through graduate school. Thank you to my grandparents, Mack and Janice, for instilling curiosity about the natural world in me, which led me to graduate school and this project. Thanks to my mom and dad for supporting me through all the setbacks over the last few years. And thank you to Yunah for your love, being my best friend, and being my biggest supporter!

Abstract

DETECTING AND PREDICTING HOT MOMENTS OF METHANE EMISSIONS FROM COASTAL WETLANDS

2024

by

Grace Pearsall

Coastal wetlands are highly productive ecosystems and can store large amounts of carbon (C). However, decomposition processes in coastal wetlands also produce and emit greenhouse gasses (GHG), such as methane (CH4) - a potent greenhouse gas that could offset C storage in the wetland soil. Often a patchwork of vegetation and open water, coastal wetlands exhibit strong biogeochemical heterogeneity, resulting in elevated CH4 flux (FCH4) at certain times and locations. These points of elevated FCH4, termed “hot spots and hot moments" (HSHM), experience biogeochemical rates so high they can disproportionally contribute to annual flux rates. Despite the broad utilization of the term HSHM, there is no standardized, statistically rigorous method for identifying HSHM and quantifying their impact on ecosystem processes. Furthermore, the conditions that trigger HSHM of FCH4 are poorly understood, and hot moments are often excluded from wetland FCH4 upscaling and predictive modeling. This study presents a comparative analysis of standard HM identification techniques to find the best HM detection method for coastal wetlands and formalize HM identification best practices. We found that using a rolling Z-score threshold to identify hot moments from eddy covariance (EC) flux

data was most suitable for coastal wetlands. Using this approach, we flagged hot moments at nine wetlands in the San Francisco Bay-San Joaquin River Delta (Bay-Delta). We then used the identified HMs to train several data-driven Random Forest (RF) models that leverage EC data to predict the occurrence of HMs. The best performing RF accurately (79%) captured HM absence/presence in the Bay-Delta region, and the relative importance of predictive environment parameters in the model shed light on the best predictors for HM. The method comparison in this study provides a best practices workflow for researchers when defining HSHM, and the RF HM model provides an upscaling methodology that could be used to predict the occurrence of HM FCH4 at sites without EC towers. Thus, the HM identification methodology and the predictive model present a valuable tool for wetland managers and restoration planners who can use the information to prioritize time and resources for mitigating and preventing these rare but high-impact emission events.

Chapter 1: Identifying Hot Moments of FCH4 at Coastal Wetlands

Introduction and Background

In the 20 years since McClain and other's seminal 2003 paper formalized the hot spots and hot moments (HSHM) concept, the HSHM framework has become ubiquitous in biogeochemical and ecological studies seeking to understand spatiotemporal extremes in ecosystem components and processes (Walter et al., 2023). The rise in recognition of the HSHM paradigm has brought attention to the problem that rare locations and events that have highly disproportionate impacts on overall biogeochemical processes in an ecosystem are often missed during sampling or discounted as outliers (Iglewicz and Hoaglin 1993; Rousseeuw and Hubert 2011). McClain et al. 2003 defined hot spots as locations that experience disproportionately high reaction rates relative to surrounding areas and hot moments (HMs) as brief points in time that show disproportionately high reaction rates relative to intervening periods. Identifying and quantifying HSHM ensures that the spatiotemporal dimensions of extreme measurements that often have an undue influence on overall ecosystem functioning and are critical for accurate modeling and upscaling approaches are not lost (Walter et al., 2023; Bernhardt et al., 2017)

Many authors have built upon McClain et al.'s original conception of HSHM, most notably Bernhardt et al. in 2017, who proposed a reframing of HSHM as Ecosystem Control Points, which moves away from the "hot or not" dichotomy and

implies there is a gradient in biogeochemical activity. Bernhardt et al. (2017) also took stock of the impact and usage of HSHM biogeochemistry and ecosystem. They found that the HSHM concept fostered a large volume of work examining the spatiotemporal aspects of rare biogeochemical activity, providing new insights into their impact on overall ecosystem functioning. However, despite the broad utilization of the term HSHM, Bernhardt et al. (2017) discovered no standardized, statistically rigorous method for identifying HSHM and quantifying their impact on ecosystem processes. Surveying the literature, Bernhardt et al. 2017 noted that the term HSHM is often invoked without any mathematical or statistical definition of what makes these times and locations "hot." Even as recently as 2023, researchers such as Walter et al. 2023 still note that a standard method for quantifying HSHM does not exist.

A standardized, statistically rigorous HSHM identification methodology is needed to allow full utilization of the HSHM framework in biogeochemistry and ecosystem science. The absence of a standard HSHM identification methodology also presents a barrier to comparative HSHM studies across ecosystems and incorporating HSHM into predictive and mechanistic models. Invoking the HSHM language without any mathematical demonstration that the points are "hot" dilutes the meaning of HSHM as instances and locations that experience extreme reaction rates and exert a disproportionate impact on overall ecosystem functioning to just another descriptive word for peaks in a dataset (Bernhardt et al., 2017). Without a best

practice approach for identifying HSHM, it is challenging to compare HSHM identified at different sites or by different researchers, which limits our ability to study the drivers and triggers of HSHM systematically and comparatively and incorporate them into predictive models. With no standard HSHM identification scheme. One's HSHM identification methodology choice introduces subjectivity into the HSHM identification, which complicates HSHM comparative analysis because the measurements flagged as HSHM by one method might not be considered HSHM by other metrics.

Ultimately, the HSHM identification method holds much weight in understanding hot phenomena, their contribution to flux, and their drivers. While studies have outlined the need for a standard, statistically rigorous definition of HSHM or presented specific methods for HSHM identification, a comprehensive comparison of the most common HSHM identification methods, to our knowledge, has not been done yet. In this study, we aim to highlight 1) the need for HSHM identification best practice methods, 2) the different results obtained with common HSHM identification methods, and 3) the importance of choosing a statistically appropriate metric.

To that end, we tested six common HSHM detection strategies with two high-resolution eddy covariance flux tower (EC) datasets recording methane flux from two wetlands in the San Francisco Bay-Sacramento-San Joaquin River Delta (Bay-Delta)

in California, USA to demonstrate the impact a given HSHM identification method can have on the final flagged HSHM. We also utilized statistical metrics to explicitly test for the presence of HM in any data distribution (Walter et al., 2023; Darrouzet-Nardi et al., 2011) and quantified the contribution to the annual flux of the HMs identified by each method. HSHM of CH4 emissions have been documented at numerous wetlands around the globe, and the stochastic nature of wetland CH4 emissions is widely appreciated (Waldo et al., 2020; Savage et al., 2014; Obregon et al., 2023; Tupek et al., 2015; Rey-Sanchez et al., 2022; Anthony & Silver, 2023). For this study, we focus solely on HMs of CH4 flux (FCH4) since our dataset has hourly temporal resolution and spans over ten years of data, making it perfectly suited for HM detection. Additionally, the statistical methods typically used to identify HSHM are usually spatiotemporally agnostic, meaning that statistical indices applied to flag HMs from intervening periods can also be applied to flag HS in an intervening matrix.

Data Source and Site Descriptions

2.1 Data Source and Pre-processing

The FCH4 data utilized in this study was collected using the EC flux method at MYB Wetland and EDN and accessed through the AmeriFlux network (Novick et

al., 2018). EC is a micro-meteorological method that directly observes gas, energy, and momentum exchanges between ecosystems and the atmosphere (Baldocchi, 2003). The MYB Wetland and EDN flux towers were installed in October 2010 and February 2018 and operated by the Biometeorology lab at UC Berkeley and the Oikawa lab at Cal State East Bay, respectively. All data published in the AmeriFlux network undergoes processing, which includes quality control and assurance (Chu et al., 2023). Since the AmeriFlux data from MYB and EDN have already undergone rigorous processing, we assume that any extreme values present in the dataset are not outliers caused by analytical or instrument errors but are real, extreme FCH4 measurements. All data were downloaded at the half-hour resolution, and to reduce noise, we aggregated half-hourly measurements into hourly measurements.

2.2 Site Descriptions

For this method comparison, we set up a case study using EC data from two hydrologically distinct wetland sites in the SF Bay-Delta, MYB Wetland and EDN Marsh. We chose these two sites to determine if and how the HM identification methods might perform differently when applied to the same flux at different sites. Because of their distinct hydrologies, restoration histories, and vegetation cover, MYB and EDN have very different greenhouse gas budgets and CH4 emission spatiotemporal patterns. Comparing the various HM identification methods'

proficiency at flagging FCH4 HM at two sites with different emission patterns will reveal discrepancies in the identification methodologies.

MYB wetland was restored from a livestock pasture into a freshwater wetland in 2010, and the current land cover is 50% open water and 50% vegetation. Managed by the California Department of Water Resources, water from the nearby river is piped in during dry summers to maintain water levels (Arias-Ortiz et al., 2022). MYB is a high-emission site with a mean FCH4 of 116.72 nmol m^{-2} hr^{-1} and a large range of FCH4 values. At varying seasons, MYB can act as a net sink of carbon (-223 ± 79 g C m-2 yr^{-1}) or a net source of methane (50 ± 5 g C-CH4 g m^{-2} yr^{-1}) (Hemes et al., 2019). Figure 1 illustrates a pronounced seasonal cycle in FCH4, with emissions peaking during the growing season around August each year, correlating with heightened plant and microbial activity associated with CH4 production.

EDN is a tidally influenced, restored polyhaline marsh in in the San Francisco Bay. Restored from salt ponds in 2008, it consists of 80% mudflats and 20% vegetated areas. Managed by the California Department of Water Resources, EDN is tidally connected to the Bay with a tidal range of 1.34 m (Arias-Ortiz et al., 2023). EDN is a lower-emission site than MYB, with a mean FCH4 of 1.80 nmol m^{-2} hr^{-1}. The site is also an efficient annual net sink for CO2 storing −387 g C-CO2 m^{-2} yr^{-1} (Shahan et al., 2022). As seen in Figure 1, the EDN FCH4 time series does not exhibit the same sinusoidal cycle in FCH4 that is seen in the MYB data. There is variability

in the FCH4 over time, but FCH4 is not consistently higher during the growing season than in the non-growing season, as seen in the MYB data.

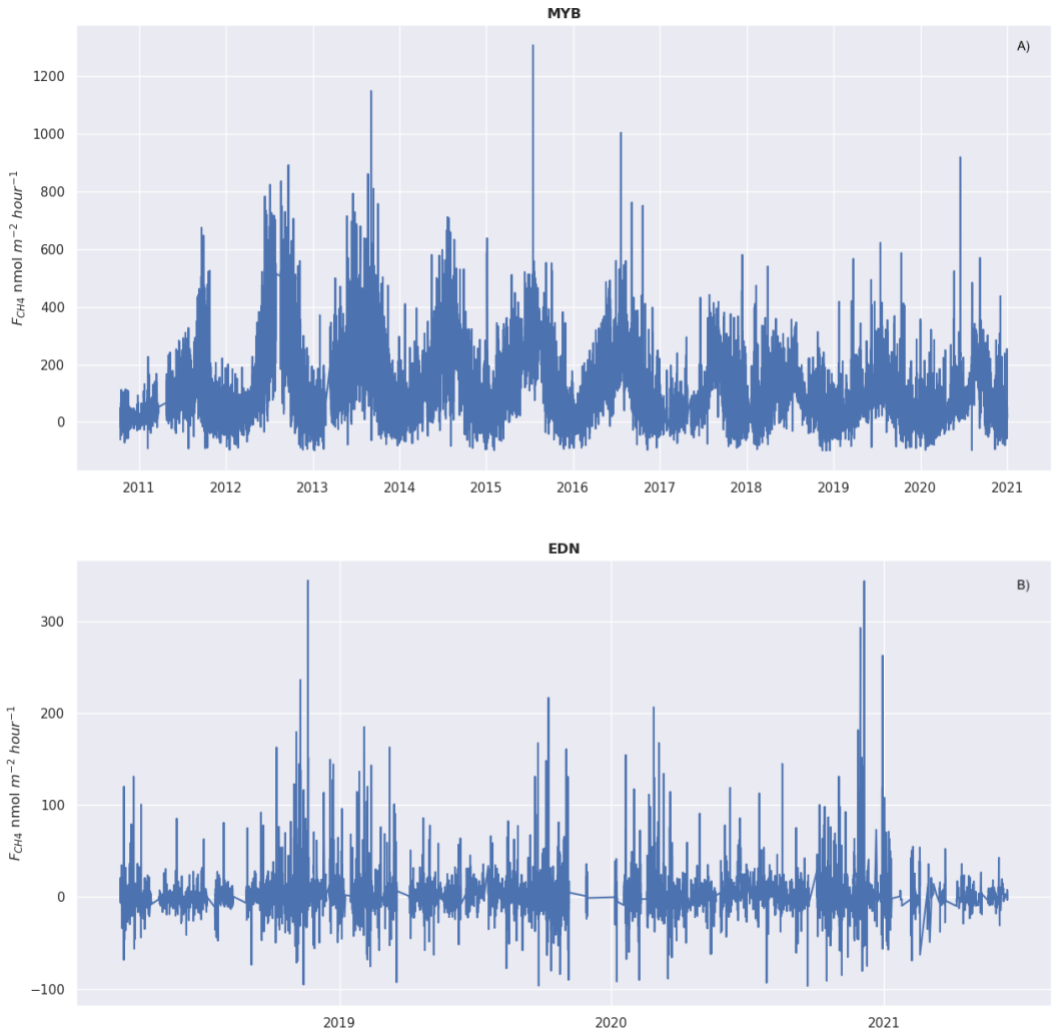


Figure 1: FCH4 time series for MYB from 2010 - 2020 (A) and EDN from 2018 - 2021 (B). Fluxes are reported on an hourly time scale.

Methods

In this systematic method comparison and review, we first assess three techniques for detecting for the presence of HM in a dataset: a skewness and kurtosis test, calculating the Control Point Influence metric, and creating Lorenz Curves for FCH4. Next, we compared six thresholds for identifying and flagging HM in a timeseries: the Z-score, Empirical Percentile, Boxplot Outlier, Reference Distribution Percentile, Rolling Z-score, and Order of Magnitude. Each of the HM identification methods use statistics indices to set a cutoff value above which, all measurements are flagged as HM. All methods tested in this study are listed in Table 1 along with the most relevant associated references. The following sections outline the methodology behind the HM presence tests, six identification methodologies we applied to the FCH4 data at MYB and EDN, and each identification method's grounding in the existing literature. We also briefly outline the methodology for and motivation behind detrending the data at MYB before applying HM identification techniques.

Table 1: HM Detection and Identification Method Summary

	Method Name	**References**
HM Presence Detection	Skewness and Kurtosis Test	Walter et al., 2023
	Control Point Influence	Arora et al., 2022
	Lorenz Curves	Saha et al. 2017
HM Identification	Z-Score Threshold	Kannenburg et al., 2020
	Empirical Percentile Threshold	Anthony and Silver et al., 2023
	Boxplot Outlier Threshold	Molodosky et al., 2012; Johnson et al., 2010
	Reference Distribution Percentile Threshold	Darrouzet-Nardi et al., 2011; Walter et al., 2023
	Rolling Z-Score Threshold	Waldo et al., 2021; Hagedorn & Bellamy, 2011; Woodrow et al., 2022
	Order of Magnitude Threshold	Vidon et al., 2010

3.1 Detrending MYB Data

Since HM are defined as events that are disproportionately offset from the mean if there are trends in the data set such as diurnal, tidal, or seasonal variability or long-term changes over the record (for example, due to climate change), calculating offsets from annual or decadal averages would result in classifying some data as HM where they should not. For example, in our data, CH4 pathways and emissions in wetlands can vary seasonally or over time since restoration since CH4 emissions are

closely linked with plant activity, primary production, microbial activity, and air and soil temperature. MYB wetland exhibits a strong seasonal signal in FCH4, where emissions are higher during the growing season when plant productivity is high, and CH4-producing microbes are most active. With higher FCH4 during the growing season (June to August), most data in that season fall above the mean FCH4. Therefore, when metrics that assess each measurement of FCH4 as the distance above the mean are used, there is a tendency to over-identify HMs in the growing season and miss elevated fluxes during the non-growing season.

For example, when we define HMs as three standard deviations (SD) away from the FCH4 mean, we almost exclusively flag HMs during the growing season when fluxes are elevated above the dataset's mean (Figure 2). After removing the seasonal signal, the HMs flagged with three SD from the de-trended mean are more consistently spread throughout the year between growing and non-growing seasons. It bears noting that there will likely always be slightly more HMs during the growing season when CH4 pathways are more active because, with increased activity, anomalously high emissions are statistically more probable. We found over-identification of HM when using the data without detrending regardless of the statistical method used to detect HM (Z-score, percentile, boxplot outlier, reference distribution cutoffs; see Supplementary Materials) and used stationary MYB data with each of these methods.

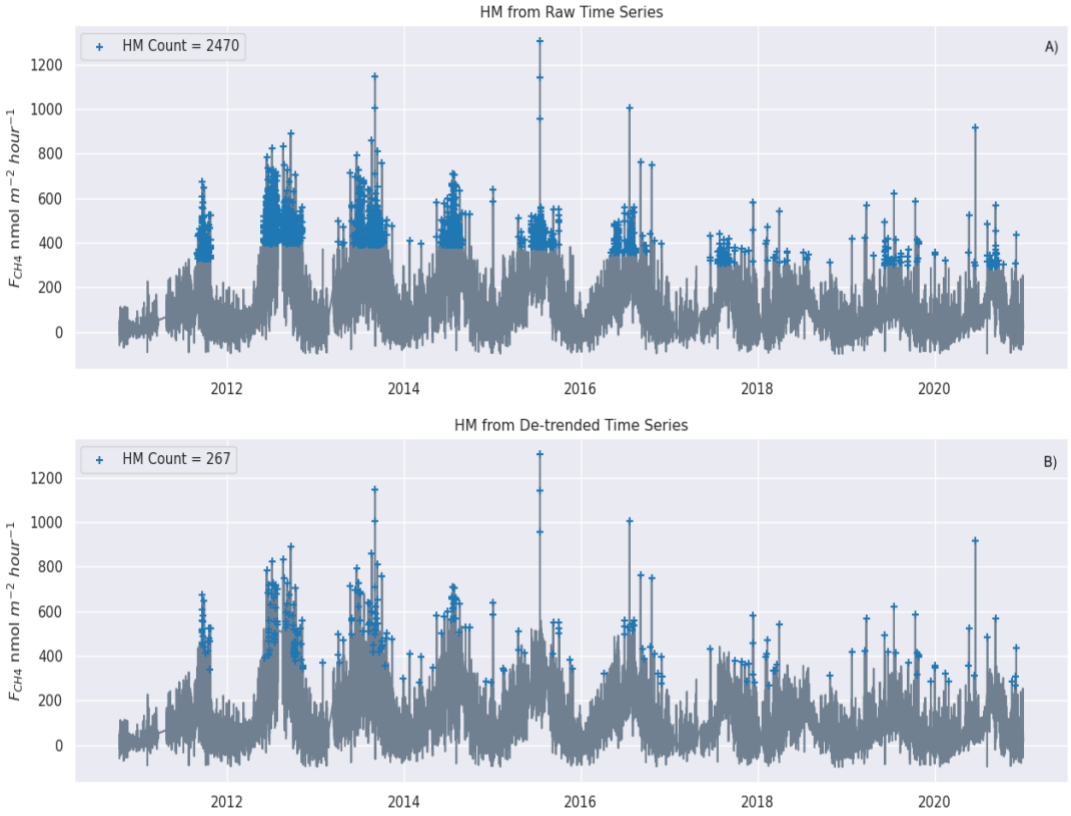


Figure 2: HM of FCH4 at MYB detected in the unaltered data (A) and in the de-trended data (B). HMs are shown denoted by blue cross marks.

In the MYB data, we used seasonal differencing, a method employed to assess the drift of statistical properties in non-stationary data that can be used to remove seasonal signals, yielding a stationary time series to detrend the record (Birkel et al.,

2014). We removed the seasonal cycle by computing the monthly mean FCH4 at MYB and subtracted the corresponding monthly mean from each FCH4 measurement. For example, we subtracted the January mean FCH4 from all January FCH4 measurements and repeated the procedure for all other months, yielding a stationary FCH4 time series. Detrending the FHC4 time series at MYB allowed us to use distribution-based HM identification methods without over-identifying HMs in the growing season when FCH4 measurements are higher than the dataset's mean. Following the detrending of the dataset, we explored approaches for testing for the presence of HM in a dataset.

3.2 Hot Moment Prescence Detection Tests

As noted by Walter et al. 2023, one of the significant limitations in current HM identification approaches is that most workflows do not explicitly test for the presence of outliers in a distribution before trying to identify HSHM. Instead, studies have historically relied on a qualitative definition of HMs and assumed that there are HMs in a dataset as long as there are distinct peaks or outliers in the timeseries (Molodsky et al., 2014; Carpenter et al., 2015; Hagedorn et al., 2011; Waldo et al., 2021; Mander et al., 2021; Obregon et al., 2023; Woodrow et al., 2022; Harms & Grimm, 2008). We present three techniques for detecting HM presence in a dataset: a skewness and kurtosis test, calculating a system's Control Point Influence, and

creating a Lorenz Curve for a system.

3.1.1 Skewness and Kurtosis Test

To detect HM presence, we can quantify the degree of 'tailedness' of the data distribution (Batt et al., 2017). Tails of the distribution describe the frequency of measurements that significantly deviate from the mean (extremes). We utilized an HM presence detection method outlined by Walter et al. (2023) that employs skewness and kurtosis statistics, which can quantify the degree and direction of tailedness in a dataset. Kurtosis measures whether the data is light-tailed or heavy-tailed compared to a normal distribution and quantifies how many measurements reside in the tails of a distribution. Skewness measures the asymmetry of a distribution and reflects whether a distribution is skewed to the left or right. Walter et al. 2023's approach utilizing both skewness and kurtosis to test for the presence of HMs presents a distinct advantage for ecosystem processes because considering kurtosis and right and left skewness allows us to search for the presence of HMs in data like FCH4 that has both negative and positive values and could have extreme observations in both tails.

The first step in this methodology is to calculate the skewness and kurtosis of the observed data and compare the empirical skewness and kurtosis to those of a reference normal distribution with the same mean and variance as the observed dataset using a parametric bootstrapping procedure. Comparing the observed

distribution to a normal reference distribution allows us to evaluate the ‘statistical rarity’ of the observed data’s skewness and compare it to a normal distribution with well-behaved tails. We used the <hotspomoments> R package (Walter et al., 2023) to perform the skewness and kurtosis tests on MYB and EDN’s FCH4 measurements (n=73,021 for MYB and n=13,681 for EDN). The reference distribution for this test was normal, with a skewness of 0 and an excess kurtosis of 0. This approach quantifies the observed data’s skewness or kurtosis relative to the skewness and kurtosis of the reference distribution as an indicator of statistically significant skewness and kurtosis. For example, a quantile value >0.95 corresponds to a kurtosis or skewness value significantly greater than expected by chance using a 1-tailed test at a type-1 error rate of 0.05, thus indicating a likely presence of HMs in the dataset.

3.1.2 Control Point Influence Metric

When dealing with ecological datasets of the scale required to identify HMs (many thousands of observations), statistically, there will always be a small number of extremes far from the mean by nature of probability in a large sample size. But, these extremes may not contribute disproportionately to an ecosystem's total flux or reaction rates. Therefore, identifying the presence of outliers in a dataset is insufficient evidence for the presence of HMs. Assuming that the presence of outliers or extreme values inherently means that a time series contains HMs fails to consider

the disproportionality component of HMs - that HMs are both statistically distinct from normal flux and contribute an outsize impact to total flux. We used the Control Point Influence (CPI) metric proposed by Arora et al. (2022) to characterize HMs' disproportionality and influence on total flux at MYB and EDN. CPI expands the ecosystem control points framework proposed by Bernhardt et al. (2017) to a new quantitative approach that compares HMs' contribution to an ecosystem's net flux by quantifying the fraction of the cumulative flux contributed by rates above the distribution's median.

$$CPI = \frac{number\ of\ measurements > median}{total\ number\ of\ measurements}$$

Arora et al. (2022) assert that CPI can be conceptualized as a biogeochemical trait that indicates the extent to which the overall biogeochemical function of an ecosystem is affected by HMs. Since CPI is characteristic of the unique system for which it was calculated, CPI can easily be compared across sites and time scales. In a system with no HMs, we would expect CPI to be around 0.5. High CPI values indicate the presence of extreme values that disproportionately influence biogeochemical function, and low CPI values indicate that extreme values are less common and less influential.

3.1.3 Lorenz Curves

Next, we used the Lorenz inequality curve (Lorenz, 1905) and the associated Gini coefficient (G) to examine the disproportionality graphically and quantitatively in FCH4 contributions. Lorenz Curves and the Gini coefficient are typically used in economics to represent a population's income distribution by plotting the cumulative income for each quantile of a population against the cumulative total income. However, several studies have employed the Lorenz-Curve and Gini coefficient in biogeochemistry to assess HSHM (Saha et al., 2017; Darrouzet-Nardi et al., 2011) and demonstrate the unequal distribution of total flux percentage across the quantiles in the data. In HSHM analysis, one can use the Lorenz Curve to graphically represent the cumulative proportion of total flux against the cumulative proportion of all flux observations. A Lorenz Curve would be a y=x line in a perfectly equal distribution (the line of equality), where each observation contributes an equal amount to the total flux. The degree of inequality in flux increases as the Lorenz Curve becomes more concave, and the gap between the Lorenz Curve and the line of equality increases. This gap between the curve and line of equality is quantified by the Gini coefficient (G).

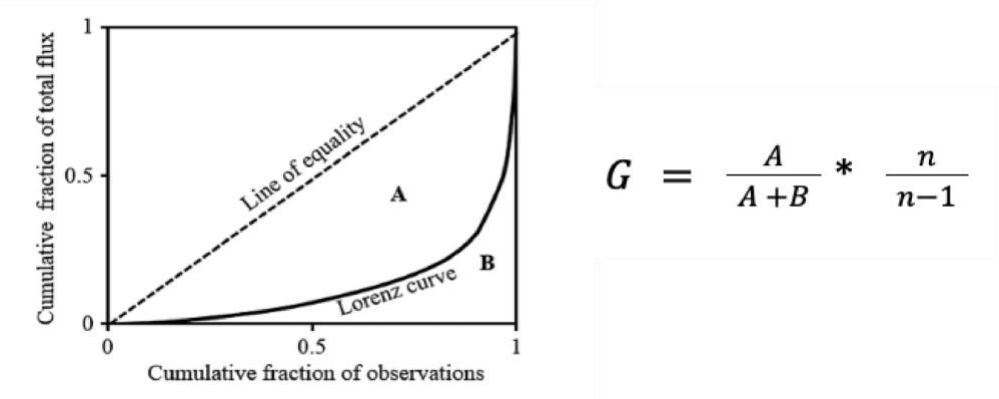


Figure 3: Example of a Lorenz Curve and Gini Coefficient equation. Figure adapted from Saha et al. (2017). G is the ratio of the area between the line of equality and the Lorenz curve (A) to the total area under the line of equality (A + B).

To give an unbiased index estimator, we multiplied G by $\frac{n}{n-1}$ where, n is the total number of observations (Pan et al., 2003; Weiner & Solbrig, 1984). The upper limit of G is one, and generally, for datasets with only positive data, the lower limit is zero. When a dataset includes negative data, obtaining a G greater than one is possible (Battisti et al., 2019), which makes G difficult to interpret. In cases with negative values, these values are typically dropped or replaced with zero before computing G. Since there were negative FCH4 at MYB Wetland and EDN, we excluded negative flux per best practice protocol (Battisti et al., 2020). Since this paper is interested in only HMs of emission and not storage, we determined that excluding negative values for Lorenz analysis is reasonable.

3.3 Hot Moment Identification Methods

After mathematically detecting the presence of HMs in the FCH4 data at MYB and EDN, we tested six common techniques for identifying and flagging HMs in a time series. We compared using Z-score, Percentile, Boxplot Outlier, Reference Distribution Percentile, Rolling Z-score, and Order of Magnitude thresholds for identifying HMs at MYB and EDN. For each method, we identified a set of HMs and assessed the contribution of the HMs to overall FHC4 at each site. For each of the six methods, we present the method's statistical foundation, previous usage of the method in HSHM studies, and how we applied the method to our data.

3.3.1 Z-Score Cutoff

HMs are inherently extreme values in a dataset. Thus, some of the most common methods for identifying HMs rely on statistical definitions and tests for the presence of outliers. Inference and analysis in ecological and biogeochemical studies typically rely on statistical testing to measure normalcy and SD to indicate variation in the processes of interest. Therefore, a measurement's distance from the mean value in a dataset can strongly indicate that the measurement is extreme or an outlier (Benhadi-Marin, 2017). One way to measure the distance of a value from the data population mean is by calculating a Z-score as

$$Z_i = \frac{x_i - \underline{x}}{s}$$

where x_i is the measurement value, $\underline{x}$ is the sample's mean, and s is the standard deviation of the sample. Z-score reports the number of SDs away from the sample mean, and each data point can be used to quantify the 'rarity' of a measurement relative to a normal distribution. Positive Measurements with Z-scores higher than ± 2 or ± 3 is considered outliers in a sample (Benhadi-Marin, 2017). Kannenburg et al. (2020) used Z-scores and EC data to identify HMs of Gross Primary Production (GPP) across different biomes by calculating the Z-score for each measurement of GPP relative to the growing season daytime mean. In this study, Kannenburg et al. set a threshold of 2 SD as an HM cutoff and flagged all GPP measurements with Z-scores > 2SD as HMs. The 2 SD cutoff has a strong grounding in literature as a threshold for climatic extremes (Anderegg et al., 2015; Huang et al., 2018; Wu et al., 2018; Kannenburg et al., 2019 & 2020; Kolus et al., 2019). A 2 SD threshold also maintains a middle ground between having a cutoff high enough to ensure the HMs flagged are extreme values and low enough that the sample size of HMs flagged is not limited. To apply the Z-score cutoff method to the FCH4 from MYB and EDN, we calculated a Z-score for every measurement at each site using Equation 3. We followed Kannenburg et al.'s (2020) precedent and defined HMs as any FCH4 measurement with a Z-score greater than 2 SD.

3.3.2 Percentile Cutoff

SD measures a sample's dispersion relative to the mean. In a distribution, SD is associated with the cumulative percentage of the data and the distribution's quantiles and percentiles. For example, in a normal distribution, the values between -2 SD and +2D comprise 95% of the data, and data above and below -2 SD and +2D comprise 2.5% of the data, respectively. Percentiles and quantiles are statistical measures that show data distribution and summarize the relative position of data within a dataset based on their magnitude, irrespective of any specific underlying probability distribution (Ialongo et al., 2019). Percentiles are quantiles that divide a distribution into 100 equal parts, and the percentile rank of a score is the percentage of scores in the distribution that are lower than that score (Everitt & Skrondal, 2010). Therefore, percentiles are another way to conceptualize the rarity of measurement in a dataset and can be used as cutoffs to define outliers and HMs.

Anthony and Silver (2021) utilize this relationship between percentile and rarity to define HMs of Nitrous Oxide and CH4 in EC data from agricultural peatland. Anthony and Silver (2021) set a higher threshold than Kannenburg et al. (2020) and defined HMs as 4 SD away from the mean or higher than the 99.9% percentile. Percentile cutoffs have also been used to define other climatic extremes, such as marine heatwaves (MHW) (Hobday et al., 2018; Giamalaki et al., 2021), where heat

waves were flagged when sea surface temperature anomalies were above the 90th percentile. In this study, we calculated empirical percentiles for the data from MYB and END using,

$$P = \frac{n}{N} * 100$$

where n is the number of data points below the data point of interest and N is the total number of data points in the data set. We tested the 90th, 95th, 97.5th, and 99.9th percentiles as HM cutoffs, and we determined that the 97.5th percentile cutoff was most appropriate for this study as it corresponds to the typical 2 SD HSHM threshold (see Supplementary Materials for other percentile test results).

3.3.3 Boxplot Outlier Cutoff

Like percentiles, quartiles divide observations in a sample into four distinct, equal intervals determined by the values of the data relative to the entire dataset. These quartiles are categorized into lower (Q1), median (Q2), and upper (Q3) quartiles, where 25%, 50%, and 75% of the data fall below the quartiles, respectively (Langford, 2006). Quartiles can be visualized with box and whisker plots that display the median, lower, and upper quartiles, and minimum and maximum values in a dataset. Quartiles summarize the central tendency and variability in a dataset, and because they represent the spread of data, quartiles are often used in methods for

outlier detection (Benhadi-Marin, 2017). A common technique is the Tukey Method, which uses the quartiles and interquartile range (IQR) to filter high and low outliers (Tukey, 1977) with the following formulas:

$$Low\ outliers\ =\ Q_1\ -1.5\times IQR$$

$$High\ outliers\ =Q_3\ +1.5\ \times IQR$$

Where $Q_1$is the first quartile, Q_3 is the third quartile, and the interquartile range (IQR), calculated by $Q_3 -\ Q_1$. This method is also called the box plot outlier method and has been utilized in several studies to identify HMs (Li et al., 2015; Molodosky et al., 2012; Johnson et al., 2010; Barnes et al., 2023; Philippe & Karume, 2019). The Tukey method relates the magnitude of each measurement to the median rather than the mean and, as such, can be used even when data is not normally distributed (Molodosky et al., 2012). The studies that use the box plot outlier method to detect HMs typically use the formula for $High\ outliers$ and refer to the result of that formula as the $Upper\ Fence$ (UF) that acts as the numerical cutoff for HMs. The $High\ outliers$ formula can be modified to differentiate degrees of outliers (mild, severe, extreme) by setting the fence a fixed distance from the IQR. Molodosky et al. (2012) and Johnson et al. (2010) calculated a mild and extreme UF to identify HMs

and asses which UF is most suitable using the following:

$$UF_{mild} = Q_3 + 1.5(IQR)$$

$$UF_{extreme} = Q_3 + 3(IQR)$$

We calculated the quartiles for each site to apply the box plot outlier method to the FCH4 data at MYB and EDN. Then, we calculated the mild and extreme UFs and flagged all measurements that fell above the mild UF and below the extreme UF as mild HMs and measurements that fell above the extreme UF as extreme HMs. We determined that the extreme UF was a more conservative and discerning HM threshold and presented the results of the extreme UF HM test here (Mild UF results can be found in Supplementary Materials).

3.3.4 Reference Distribution Cutoff

The methods mentioned above all compared flux measurements to the empirical distribution of the data and identified outliers and HMs relative to the median, mean, or quartile values calculated from the data. The following approach compared the empirical flux data to a reference distribution and flagged measurements as HMs if they were more extreme than we would expect to occur in a reference distribution (Batt et al., 2017). Two examples of this approach can be found in Walter et al. 2023 and Darrouzet-Nardi et al. 2011, who used a Normal distribution

and Student's T distribution as references to identify HMs, respectively. For this study, we opted to test Walter et al.'s (2023) normal distribution reference methodology as a case study of this reference distribution approach. This normal distribution reference approach extends the same logic as the skewness and kurtosis HM presence test proposed by Walter et al. (2023), assuming that since HMs are inherently extreme and datasets including HMs will have higher skewness or kurtosis than datasets without HMs. Therefore, when we compare a dataset that includes HMs to an analogous reference distribution, the HMs will stand out as exceptionally more extreme than in a reference distribution. In a normal distribution, <4.6% of observations are more than 2 SD away from the mean, and <0.3% of measurements are more than 3 SD away from the mean. As noted by Walter et al. 2023, since extremes in a normal distribution are so rare, we can assume that extremes in a normal distribution are proportionally too rare to impact overall ecosystem functioning disproportionately and are not HMs (Walter et al., 2023). This assumption allows us to compare a reference normal distribution to our empirical dataset with high skewness and kurtosis, which proportionally has more extreme values that we can classify as HMs. To use the reference distribution to flag HMs, one must designate a percentile cutoff for the reference distribution and flag all measurements from the empirical dataset above that percentile as an HM (Walter et al., 2023).

We applied this normal distribution reference method to our FCH4 using Walter et al.'s 2023 R package <hotspomoments>. The reference distribution for both sites was normally distributed, and we determined the Normal distribution's percentiles and flagged any measurements in the empirical data whose percentile is greater than the Normal distribution's cutoff. We found that using the 97.5th percentile rather than the 95th percentile that Walter et al. 2023 employed provided a more conservative HM threshold. This approach differs from the percentile cutoff method mentioned in section 3.3 because the cutoff measurement corresponds to the 97.5th percentile in the normal reference distribution and not the measurement corresponding to the 97.5th percentile in the empirical data.

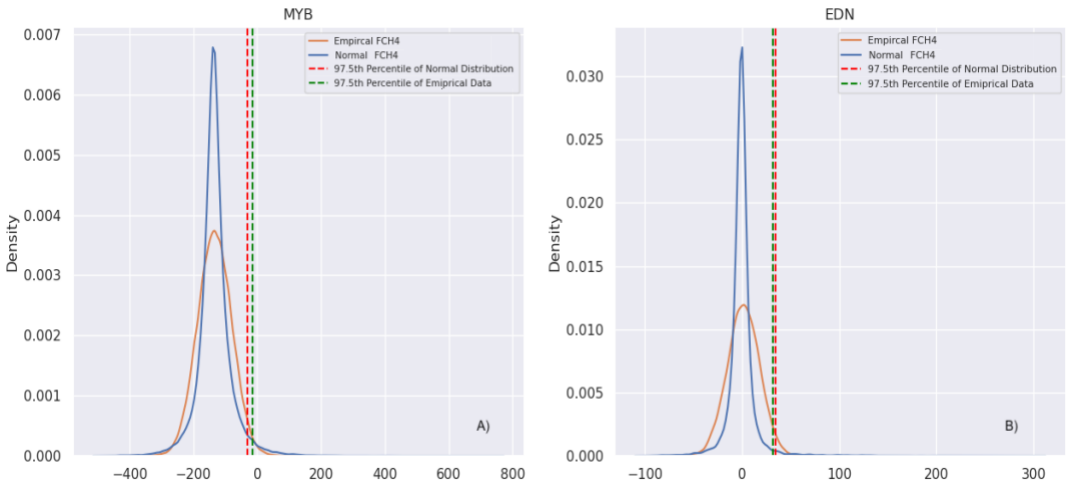


Figure 4: Empirical and normal distribution of FCH4 for MYB (A) and EDN (B). Empirical and reference normal distribution 97.5th percentile HM cutoffs are shown

as dotted lines.

3.3.5 Rolling Z-Score Cutoff

The most basic definition of HMs frames them as short flux events distinct from a given system's baseline, average flux. Many studies use this concept to qualitatively define HMs as any time fluxes are ‘elevated’ from baseline rates (Waldo et al., 2021; Hagedorn & Bellamy, 2011; Woodrow et al., 2022), describing HMs as ‘flux peaks’ (Obregon et al., 2023) or ‘remarkably high flux values’ (Mander et al., 2021). In this study, we synthesize this baseline flux concept into a quantitative approach, calculating a baseline flux time series and comparing each FCH4 measurement to its corresponding measurement in the baseline flux. We then flagged all statistically distinct measurements (see below) from baseline flux as HMs. For this methodology, we borrowed concepts of extreme climatological event identification techniques that compare events to seasonal and climatic means and the Hampel Filter outlier identification strategy. Constructing baseline timeseries or temperature, sea surface anomalies, and precipitation has been used to identify extreme temperature events, marine heat waves, and extreme precipitation events by comparing anomalously high events to the constructed baseline (Hobday et al., 2018). The Hampel Filter uses the median absolute deviation and a moving window to detect outliers (Hampel, 1971) by comparing each measurement to its neighbors in the

moving window and is often employed in ecological studies.

In the methodology described here, we defined baseline flux using a moving window to smooth the hourly FCH4 time series into a seasonal time series (experiments with weekly and monthly rolling averages shown in Supplementary Material). We then used another moving window of the same size, informed by the Hampel filter, to calculate the moving Z-score for each measurement and flagged every measurement with a moving Z-score above three SD as an HM. Applying a moving average to a time series yields a smoothed curve that displays the dominant signal in a time series - which we can conceptualize as the baseline FCH4. The moving average is calculated for a given window of time using the following Equation 5:

$$\underline{x_i} = \frac{1}{w} \sum_{j=i-w}^{i-1} x_j$$

where $\frac{1}{w}$ is the window period and x_j is the measurement. Then, we calculated the moving standard deviation for each window using Equation 6,

$$S_i = \sqrt{\frac{1}{w} \sum_{j=i-w}^{i-1} (x_j - \underline{x_i})^2}$$

where $\frac{1}{w}$ is the window period, x_j is the measurement, and $\underline{x_i}$ is the moving average. Determining the SD for each measurement relative to the associated moving average allowed us to calculate a rolling Z-score. The rolling Z-score applies the same logic as a standard Z-score and assesses the difference between each measurement and the mean of the same window. We can calculate the moving Z score for each FCH4 measurement with Equation 7,

$$Z(x_i) = \frac{x_j - \underline{x_i}}{S_i}$$

where x_j is the current measurement, $\underline{x_i}$ is the current window's moving average, and S_i is the current window's standard deviation. This calculation transforms the distance from the moving average to a Z-score with SDs as the unit. We can then take the Z score for each FCH4 measurement, set a threshold, and say that any measurement X number of standard deviations above the moving average is an HM. The moving Z-score methodology is often used in finance and stock trading (Mare & Moreira, 2017; IBM DocumentationI). The most common application of rolling Z-scores in stock trading is where analysts use Z-scores to identify points anomalously above or below a rolling average of stock prices and thus determine when a stock

might be overbought or oversold. The application of the rolling Z-score methodology to HMs is analogous to the trading application, and we can use the rolling Z-score to identify FCH4 measurements that are anomalously high above the rolling average of FCH4. Our seasonal smoothing window period was three months or 2220 hours, and we computed the moving Z-score for each measurement in the timeseries using the same window size. The rolling averages and Z-scores were computed in Python. Following statistical best practices and experimental HM threshold testing, we set our Z-score threshold for HM identification at 3 SD above the rolling average.

3.3.6 Order of Magnitude Cutoff

Vidon et al. (2010) presented another quantitative conceptualization of defining HM as measurements distinct from baseline flux, proposing that HSHM are any site or time period where rates are one order of magnitude greater than the surrounding area or time interval. For this study, we interpreted Vidon et al.'s (2010) definition to mean that any measurement one order of magnitude greater than its corresponding seasonal flux measurement was an HM. We used the rolling seasonal baseline we created at MYB and EDN to apply this technique to our data and set the threshold as one order of magnitude away from this baseline flux. We calculated the equation below, where $FCH4_i$ is each measurement and $\underline{x_i}$ is the seasonal mean of FCH4.

$$FCH4_i \geq \underline{x_i} * 10\,,$$

Results

4.1 Hot Moment Presence Detection Tests

Using Walter et al.'s (2023) skewness and kurtosis HM presence test, we found that the MYB FCH4 dataset had a skewness of 1.40, a kurtosis of 2.55, and a quantile comparison statistic of 0.99 for both metrics. The EDN FCH4 data had a skewness of 5.23, a kurtosis of 71.97, and a quantile comparison statistic of 0.99 for the skewness and kurtosis test. With a comparison quantile of 0.99 for FCH4 at MYB and EDN, we can conclude there is statistical evidence for HMs at these sites.

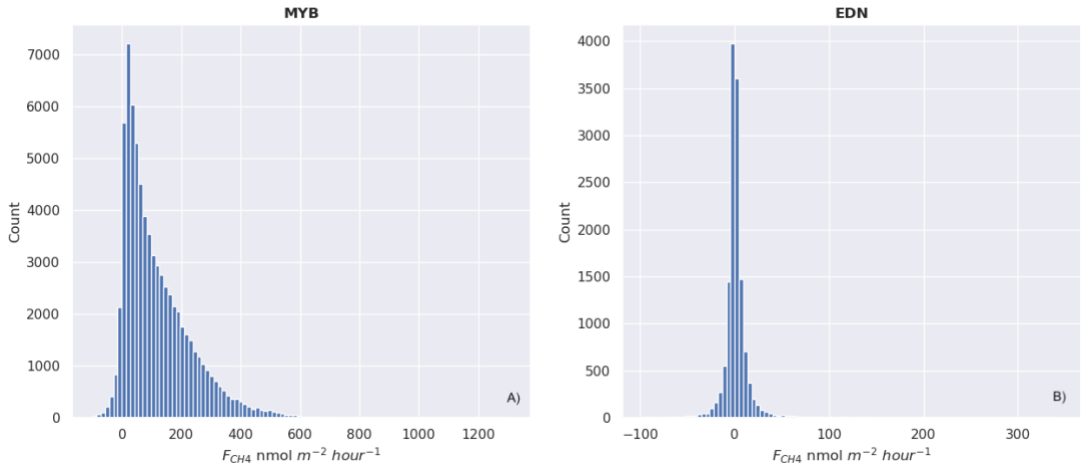


Figure 5: Histograms for FCH4 at MYB and EDN. Displays the distribution and tailedness of each dataset. The high skewness and kurtosis can be seen in MYB and

EDN. HMs fall in the right tails of each distribution.

We calculated the CPI and created Lorenz Curves for each site to test if the extreme values disproportionately impact overall FCH4. We found that at MYB, the CPI was 0.868; at EDN, the CPI was 2.65. The > 0.5 CPIs for MYB and EDN indicate that the flux at the sites is characterized by a significant number of measurements above the dataset's median that drive total flux, confirming the presence of HMs in the data. The Lorenz Curves and G generated for MYB and EDN also confirmed the presence of HMs in the data that disproportionately contribute to overall flux rates. The Gini Index for MYB was 0.47, and the Lorenz Curve was concave relative to the line of equality, which both indicate that a small portion of the fluxes in this dataset contribute a significant portion of the total FCH4 at MYB. The Lorenz Curve for EDN bows away from the line of equality even more than the MYB Curve and had a G of 0.64. These characteristics of the Lorenz Curve indicate substantial temporal inequality in measurements at MYB and EDN, where a small fraction of fluxes have an outsize influence on overall FCH4. The Lorenz Curves for EDN and MYB further indicate the presence of HMs of FCH4.

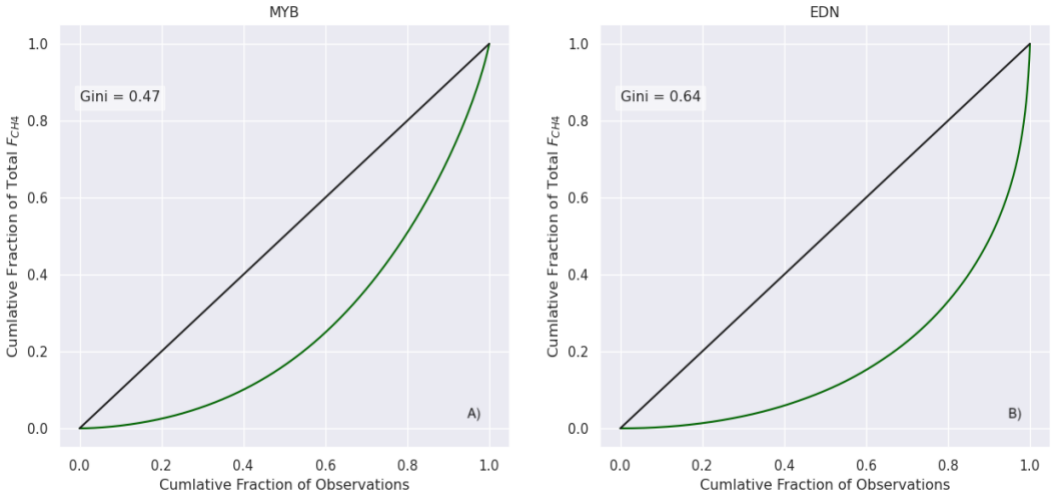


Figure 6: Lorenz Curves and Gini Coefficients for MYB (A) and END (B).

4.2 Hot Moment Identification Methods Results

4.2.1 Z-Score Thresholds

At MYB, we found 267 HMs with the Z-score cutoff. Note that the HMs were flagged with the de-trended time series but plotted on the unaltered time series in Figure 7. These HMs comprised 0.365% of total measurements but contributed 1.637% of total FCH4 from 2010-2020. We further examined the influence of stationary and non-stationary HMs on overall flux by computing the annual and total FCH4 with and without HMs and evaluating the percent change between the means with and without HMs. We found at MYB the cumulative FCH4 from 2010-2020 with HM was 8.523 x 10^6 nmol CH_4 m^{-2} hr^{-1} and without HM was 8.383 x 10^6 nmol

CH_4 m^{-2} hr^{-1}. The annual mean FCH4 at MYB with HM was 116.72 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean without HM was 115.23 nmol CH_4 m^{-2} hr^{-1}.

Using the two Z-score cutoff at EDN, we flagged 298 HMs of FCH4. Looking at the distribution of HMs in the timeseries, they occur throughout the year with no apparent bias for the growing or non-growing seasons. We found that the 298 HMs at EDN comprised 2.17% of the total measurements but 91.53% of the total FCH4, a strikingly high disproportionality between HM occurrence and flux influence. The cumulative FCH4 at EDN with HMs was 2.4639×10^4 nmol CH_4 m^{-2} $hr^{-1,}$ and without HM was 0.3746 x 10^4 nmol CH_4 m^{-2} hr^{-1} without HMs. The annual mean FCH4 with HMs was 1.80 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean without HMs at EDN was 0.156 nmol CH_4 m^{-2} hr^{-1}. Further analysis of mean FCH4 with and without HM broken down by year for MYB and EDN is available in Supplementary Materials.

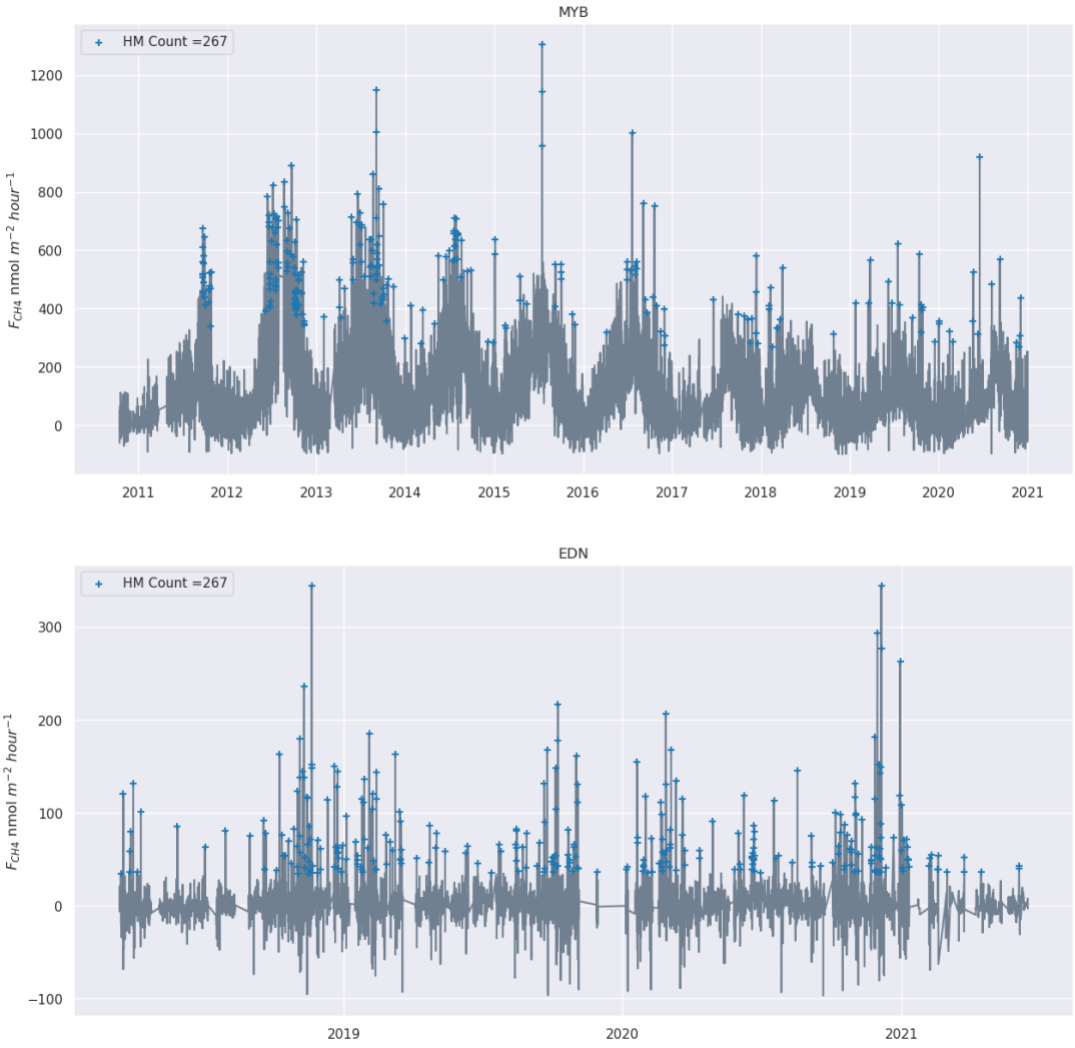


Figure 7: HMs identified with classic Z-score technique for MYB (A) and EDN (B). Z-Score cutoff was 2 SD.

4.2.2 Percentile Cutoffs

With the 97.5th percentile HM threshold, we flagged 1826 HMs at MYB.

Notably, with the percentile cutoff method, the raw and de-trended timeseries at

MYB yielded the same number of HMs as their non-stationary counterparts because percentiles describe how many observations in a dataset fall below a given measurement. Therefore, since the same number of measurements are made in the stationary and non-stationary datasets, when we flag the 97.5th percentile as a cutoff, for example, we will always flag 2.5% of the total observations as HMs (although these will not always be the same measurements flagged - see Supplementary Materials). With the 97.5th percentile, we found that HMs comprised 2.5% of all measurements and contributed 8.45% of total FCH4. The cumulative FCH4 at MYB with and without these HMs were 8.52 x 10^6 nmol CH_4 m^{-2} hr^{-1} and 7.62 x 10^6 nmol CH_4 m^{-2} hr^{-1}. The mean annual FCH4 at MYB with HMs was 116.72 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean without HMs is 109.60 nmol CH_4 m^{-2} hr^{-1}.

With the 97.5th percentile HM threshold, we flagged 343 HMs of FCH4 at EDN. The cumulative FCH4 at EDN with these HMs was 2.46 × 10^4 nmol CH_4 m^{-2} $hr^{-1,}$ and without the cumulative FCH4 was 0.1168 × 10^4 nmol CH_4 m^{-2} hr^{-1}. The annual mean FCH4 at EDN with the percentile cutoff HMs was 1.801 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean FCH4 was 0.044 nmol CH_4 m^{-2} hr^{-1}. The HMs flagged with the 97.5th percentile cutoff were 2.5% of the measurements and 97.6% of the total FCH4.

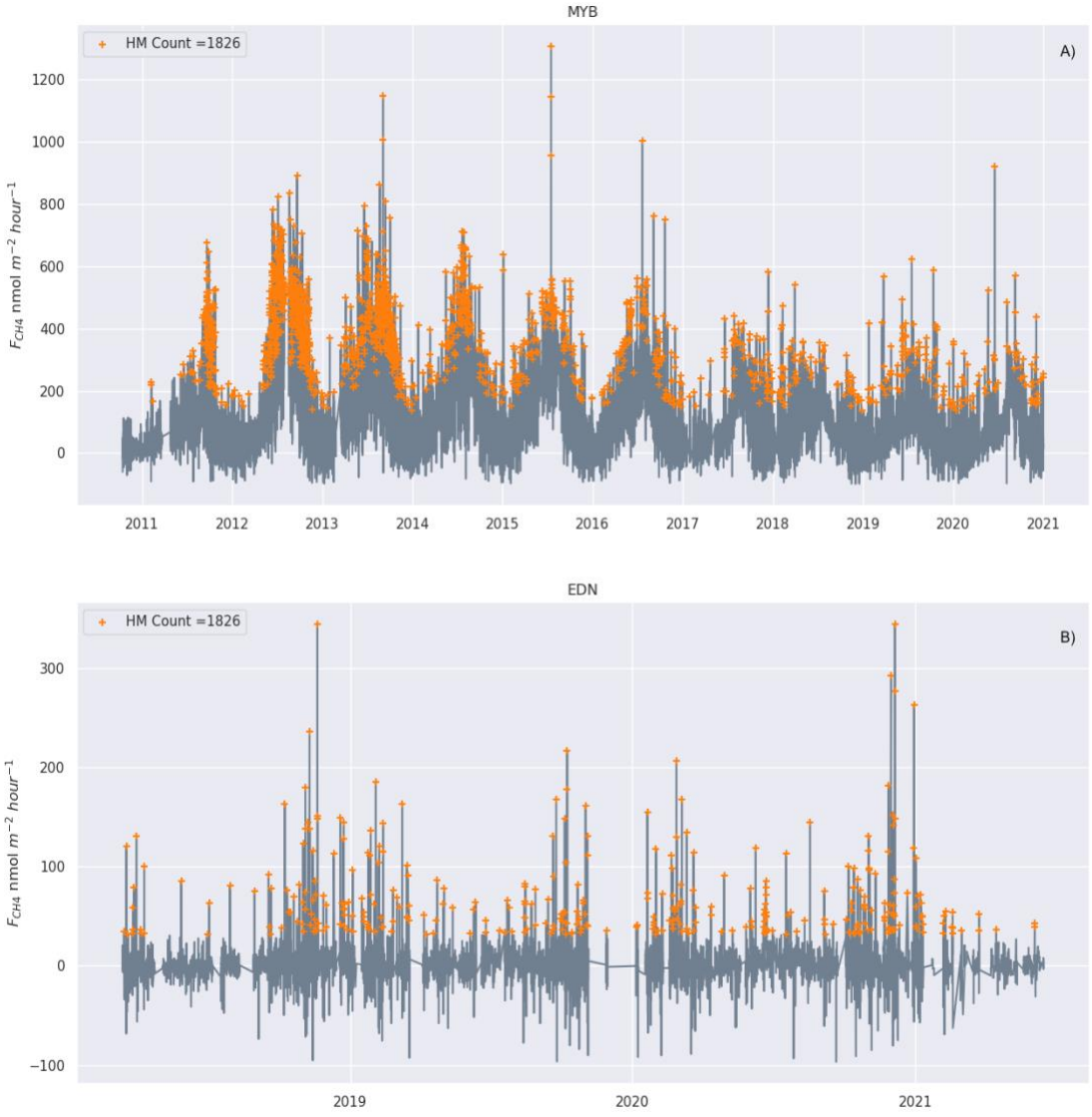


Figure 8: HMs identified using the 97.5th percentile cutoff at MYB (A) and EDN (B).

4.2.3 Boxplot Outlier Threshold

The boxplot distributions of FCH4 for MYB and EDN are shown in Figure 9, along with the extreme and mild UF HM thresholds for each site. At MYB, using the extreme UF, we flagged 1006 HMs, which comprised 1.377% of total measurements and 5.09% of total FCH4. The cumulative FCH4 at MYB with the extreme UF HMs was 8.52 x 10^6 nmol CH_4 m^{-2} hr^{-1} and without these HMs was 7.25 x 10^6 nmol CH_4 m^{-2} hr^{-1}. The mean annual FCH4 at MYB with HMs was 116.72 nmol CH_4 m^{-2} hr^{-1} and the annual mean without HMs is 96.29 nmol CH_4 m^{-2} hr^{-1}. When we applied the extreme UF at EDN, we detected 436 HMs of FCH4. These 436 HMs flagged by the extreme UF comprised 3.19% of the total measurements and 108.6% of the total FCH4. The cumulative FCH4 at EDN with these HMs was 2.46 × 10^4 nmol CH_4 m^{-2} $hr^{-1,}$ and without the cumulative FCH4 was 0.1043 × 10^4 nmol CH_4 m^{-2} hr^{-1}. The annual mean FCH4 at EDN with the extreme UF HMs was 1.801 nmol CH_4 m^{-2} hr^{-1}, and the annual mean FCH4 was -0.1605 nmol CH_4 m^{-2} hr^{-1}. Notably, the annual mean FCH4 without extreme UF HMs is negative because of the prevalence of negative FCH4 values in the EDN timeseries. Therefore, when we calculate the cumulative emissions from EDN, the total without HMs can be negative if the HMs make up most or all the emission events at the site.

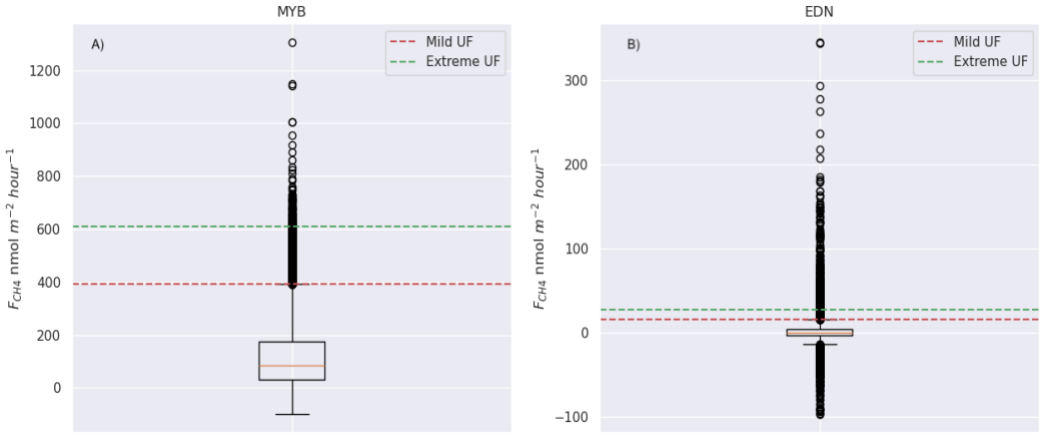


Figure 9: Boxplot distributions for MYB (A) and EDN (B). Lower whiskers denote Q1, upper whiskers denote Q3, the orange lines denote the median, and the dotted lines denote the mild and extreme UFs for HM identification.

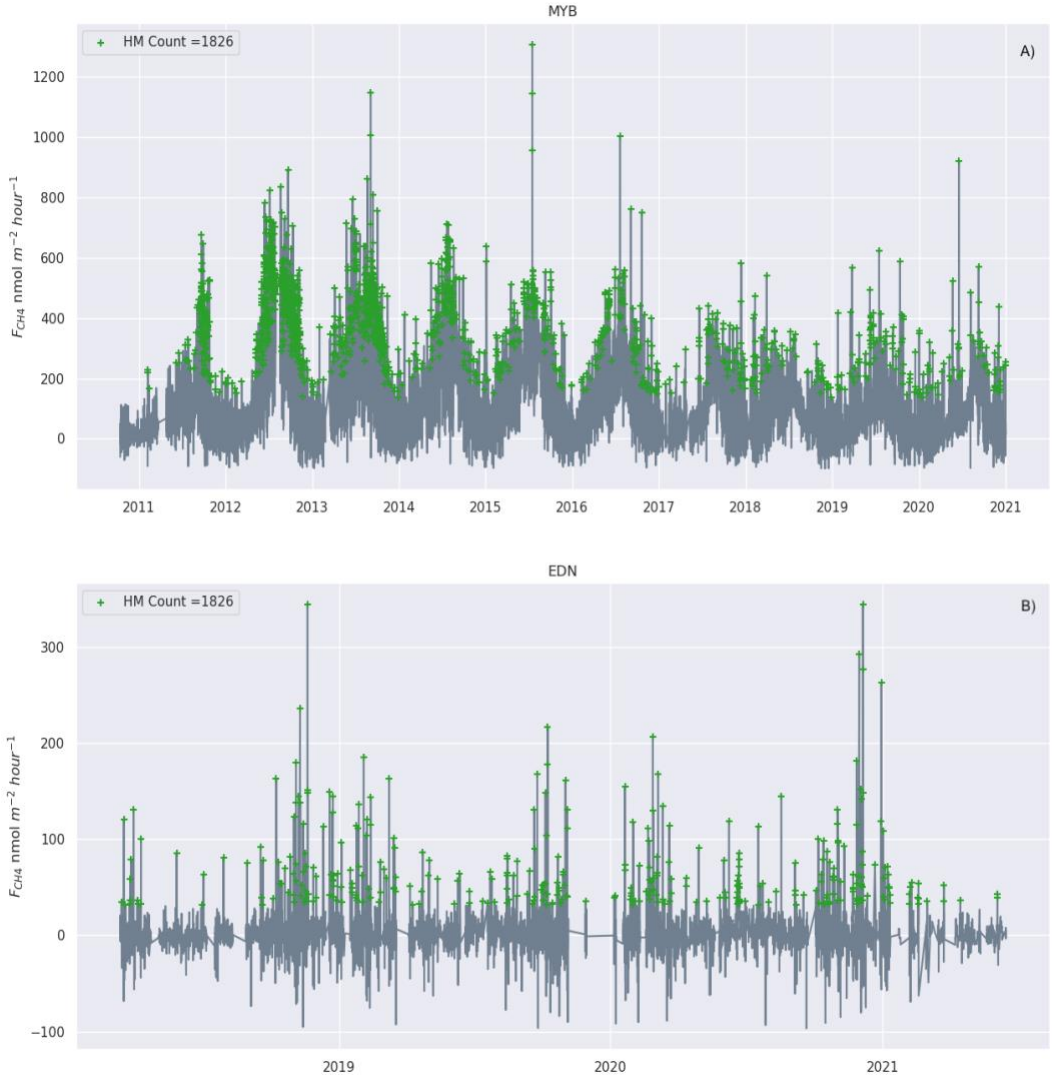


Figure 10: HMs identified at MYB (A) and EDN (B) using the boxplot outlier technique with the extreme UF cutoff.

4.2.4 Reference Distribution Percentile Threshold

For the reference distribution percentile threshold, we used a normal distribution as the surrogate distribution and set the percentile threshold at the 97.5th

percentile. At MYB, we flagged 2406 HM using the reference distribution comparison method. These HM comprised 3.29% of the total 10.35% of the total FCH4 at MYB. The cumulative FCH4 at MYB with the extreme UF HMs was 8.52 x 10^6 nmol CH_4 m^{-2} $hr^{-1,}$ and without these HMs, cumulative FHC4 was 6.93 x 10^6 nmol CH_4 m^{-2} hr^{-1}. The mean annual FCH4 at MYB with HMs was 116.72 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean without HMs was 87.89 nmol CH_4 m^{-2} hr^{-1}.

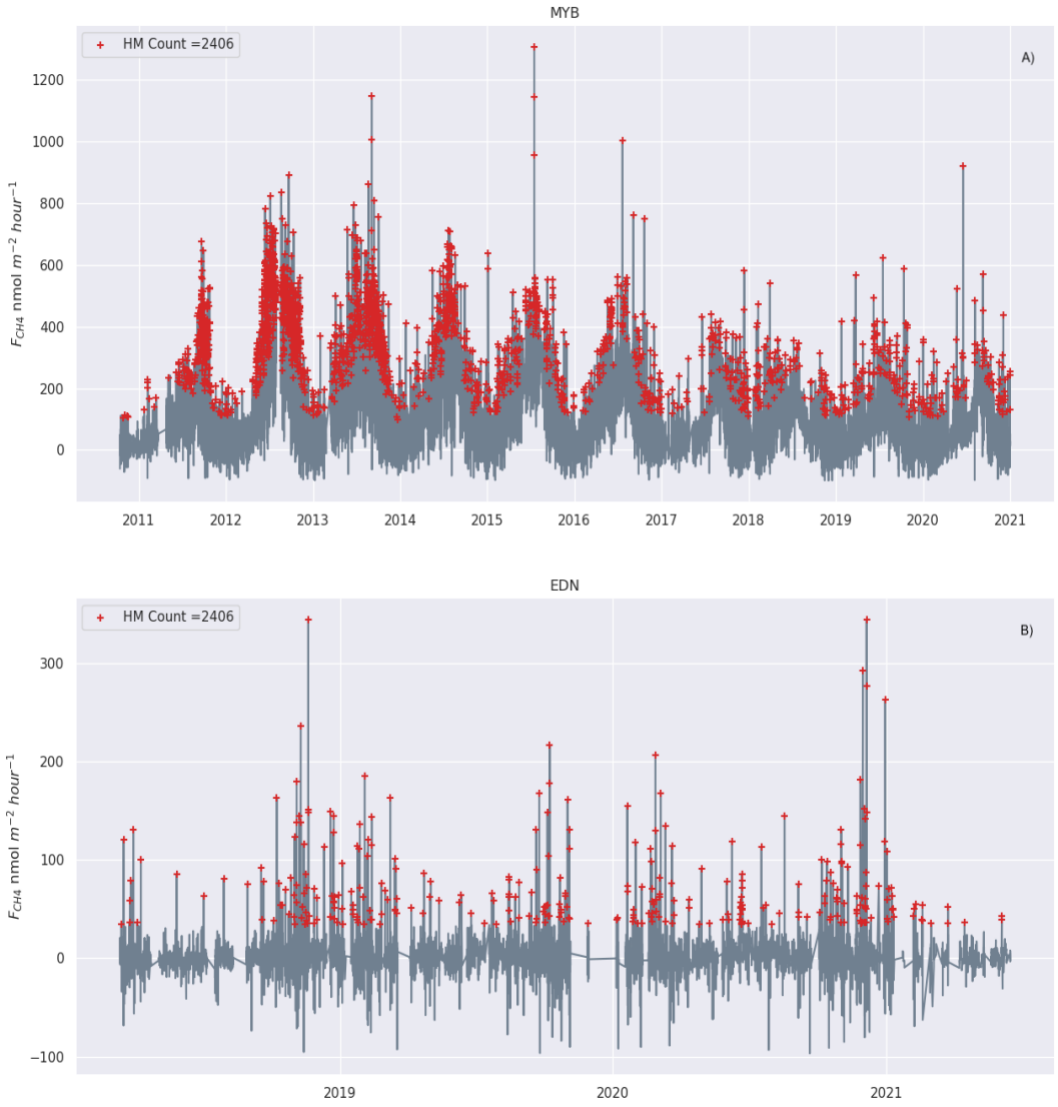


Figure 11: HMs identified using the 97.5th percentile cutoff from reference normal distributions at MYB and EDN.

Using the 97.5th reference percentile cutoff at EDN, we flagged 329 HMs of FCH4. The HMs flagged with the 97.5th reference distribution cutoff comprised

2.26% of the total measurements and 93.08% of the total FCH4. The cumulative FCH4 at EDN with these HMs was 2.46×10^4 nmol CH_4 m^{-2} $hr^{-1,}$ and without the cumulative FCH4 was 0.1704×10^4 nmol CH_4 m^{-2} hr^{-1}. The annual mean FCH4 at EDN with the extreme UF HMs was 1.801 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean FCH4 was 0.1274 nmol CH_4 m^{-2} hr^{-1}.

4.2.5 Rolling Z-Score Threshold

When we applied the seasonal rolling average as a baseline and a three Z-score cutoff, we flagged 502 HMs at MYB. It is apparent in the seasonal time series that the three standard deviation rolling Z-score was more conservative in flagging HMs than the two Z-score thresholds. When we computed the influence of the seasonal HMs flagged with a three Z-score cutoff, we found that the HMs comprised 0.9% of total measurements and 2.23% of total FCH4 at MYB. The cumulative FCH4 at MYB with the extreme UF HMs was 8.52×10^6 nmol CH_4 m^{-2} $hr^{-1,}$ and without these HMs was 8.35×10^6 nmol CH_4 m^{-2} hr^{-1}. The mean annual FCH4 at MYB with HMs was 116.72 nmol CH_4 m^{-2} hr^{-1} and the annual mean without HMs is 112 nmol CH_4 m^{-2} hr^{-1}.

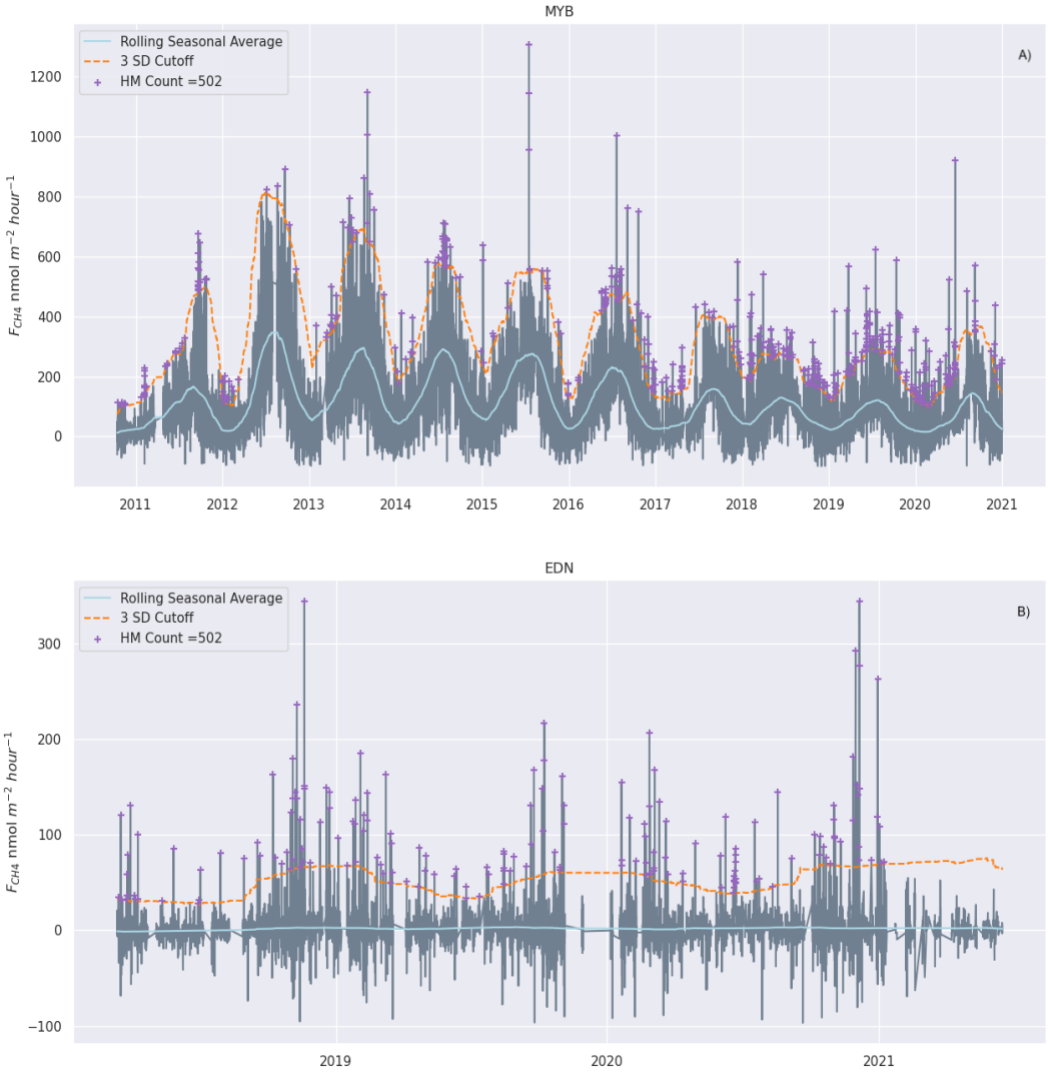


Figure 12: HMs identified with the rolling Z-score technique at MYB (A) and EDN (B). The seasonal moving average is shown in blue and the moving 3 SD cutoff is shown in orange.

Because there is no strong seasonal signal at EDN, the smoothed FCH4 time series at EDN does not exhibit seasonal cycling like the smooth time series at MYB.

Rather than smoothing the FCH4 time series and drawing out the seasonal curves, the rolling time series at EDN smooth out the extreme variability in FCH4 and yield a baseline flux that hovers around zero. When we used the seasonal rolling average as a baseline at EDN, we flagged 171 HMs using the three Z-score thresholds. Quantifying the influence of the monthly and seasonal HMs on total flux, we found that the seasonal HMs identified with a three Z-score cutoff comprised 1.25% of all measurements and 67.54% of total FCH4 at EDN. The cumulative FCH4 at EDN with these HMs was 2.46×10^4 nmol CH_4 m^{-2} $hr^{-1,}$ and without the cumulative FCH4 was 0.0846×10^4 nmol CH_4 m^{-2} hr^{-1}. The annual mean FCH4 at EDN with the extreme UF HMs was 1.801 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean FCH4 was 0.1155 nmol CH_4 m^{-2} hr^{-1}.

4.2.6 Order of Magnitude Threshold

Unlike the previous five methods, the order of magnitude cutoff approach had vastly different results at each site. As seen in Figure 13, at MYB, the cutoff was very high and only flagged 23 hot moments, which comprised 0.03% of total measurement and 0.1% of cumulative FCH4. The cumulative FCH4 at MYB with the order of magnitude HMs was 8.52 x 10^6 nmol CH4 $m^{-}2$ hr^{-1}, and without these HMs was 8.51 x 10^6 nmol CH4 m^{-2} hr^{-1}. The mean annual FCH4 at MYB with HMs was 116.72 nmol CH4 m^{-2} hr^{-1}, and the annual mean without HMs was 116.66 nmol CH4 m^{-2} hr^{-1}.

Looking at the shape of the moving cutoff values in Figure 13, we can see that ten-times-the seasonal mean produces a curve that dramatically exaggerates the seasonal signal. As a result, no measurements during the growing season come close to crossing this threshold, and HM was only flagged in the non-growing season, where the cutoff curve is lower, and some measurements can cross it.

Figure 13: HMs identified with the one order of magnitude cutoff at MYB (A) and EDN (B). The seasonal moving average is shown in blue, and the cutoff is ten times the seasonal average cutoff, which is shown in pink.

The order of magnitude approach at EDN showed the opposite results of MYB. At MYB, the order of magnitude threshold was far too high, but at EDN,

because the seasonal signal is so muted and there are negative values, the order of magnitude cutoff proved to be far too low. At the beginning of the time series at EDN in May of 2015, the seasonal baseline was around -0.9 nmol CH4 m-2 hr-1. This negative portion of the time series strongly interferes with the order of magnitude threshold because ten times the negative flux makes the cutoff value around -9 nmol CH4 m-2 hr-1nmol, and with a cutoff this low, all emissions are flagged as HM. Even when the seasonal baseline becomes positive again, the FCH4 values are so small that when multiplied by 10 to create the cutoff curve, the cutoff thresholds are still too low. This method flagged many hot moments that appear to be part of baseline flux. With the order of magnitude technique, we flagged 2126 HMs, which comprised 15.5% of the total measurement and 128% of the total FCH4. The cumulative FCH4 at EDN with these HMs was 2.46×10^4 nmol CH_4 m^{-2} $hr^{-1,}$ and without the cumulative FCH4 was -7.49×10^4 nmol CH_4 m^{-2} hr^{-1}. The annual mean FCH4 at EDN with the extreme UF HMs was 1.801 nmol CH_4 m^{-2} $hr^{-1,}$ and the annual mean FCH4 was -0.61 nmol CH_4 m^{-2} hr^{-1}

Table 2: HMs Percent Contribution to FCH4 vs Percent of Total Measurements

MYB (2010-2020)			
	HM Count	% of Total Measurements	% of Total FCH4
2 Z-Score Cutoff	267	0.367	1.64
97.5th Percentile Cutoff	1826	2.50	8.45
Extreme UF Boxplot Cutoff	1006	1.38	5.09
Rolling 3 Z-Score Cutoff	502	0.86	2.02
Reference Distribution 97.5th Cutoff	2406	3.29	10.34
Order of Magnitude Cutoff	23	0.033	0.100
EDN (2018-2021)			
2 Z-Score Cutoff	265	1.94	84.76
97.5th Percentile Cutoff	343	2.51	95.26
Extreme UF Boxplot Cutoff	347	2.54	95.77
Rolling 3 Z-Score Cutoff	171	1.37	65.66
Reference Distribution 97.5th Cutoff	290	2.12	88.4
Order of Magnitude Cutoff	2126	15.54	128.6

4.3 Comparing Hot Moment Identification Technique Performance

Each of the six HM identification techniques analyzed in this study presents a unique conceptualization of HMs and has benefits and drawbacks. Although these methods provide statistical metrics we can use to define HMs, they all include a degree of subjectivity because the user always needs to determine the appropriate cutoff value for HMs. Here we compare and rank the performance of six methods. For the data at MYB and EDN, there is no ‘true’ hot moment validation dataset that

we can use to assess the accuracy of our flagged HMs. However, we can visually inspect a time series and manually note any anomalously high fluxes and peaks (Waldo et al., 2021; Hagedorn & Bellamy, 2011; Woodrow et al., 2022) as probable HM. For smaller data sets, this method can successfully catch all HM. However, with large, high-resolution datasets that are often required to capture HMs, visual inspection is time-intensive and unfeasible. The statistical detection methods analyzed here present an 'automated' alternative that can be used with big data applied systematically across different sites and eliminates "operator" subjectivity. We assessed the performance of each method by comparing the statistically detected HMs against the HMs we visually flagged and determining if the flagged HM 1) represents distinct elevated flux events relative to intervening periods and 2) contributes disproportionately to overall site flux, and thus consistent with the conceptual definition of HSHM. To illustrate the differences between methods more clearly, we present in Figures 15 and 16 the HM flagged in one year of data that we felt best emphasized the differences between each detection method (2013 for MYB and 2020 for EDN) rather than looking at all site years of data where it is visually difficult to see all the HM. A comparison of the HMs in the complete time series for MYB and EDN is shown in Supplementary Materials and our interactive web app hosted on Heroku: https://www.heroku.com/.

Figure 14. Probable HM manually tagged at MYB and EDN. We tagged 35 probable HM at MYB and 46 probable HM at EDN.

MYB 2013			
	% of total Measurements	% of total FCH4	HM Count
2 Z-Score Cutoff	0.73	2.50	52
97.5th Percentile Cutoff	6.03	14.6	427
Extreme UF Boxplot Cutoff	3.15	8.47	223
Rolling 3 Z-Score Cutoff	0.40	1.47	30
Reference 97.5th Cutoff	6.70	16.4	502
Order of Magnitude Cutoff	0	0	0
EDN 2020			
2 Z-Score Cutoff	2.62	88.0	121
97.5th Percentile Cutoff	3.20	101.6	142
Extreme UF Boxplot Cutoff	3.78	103.7	175
Rolling 3 Z-Score Cutoff	1.49	65.7	69
Reference 97.5th Cutoff	2.77	90.2	128
Order of Magnitude Cutoff	5.99	122.7	277

Table 3: Single Year HM Flux Contribution vs Frequency.

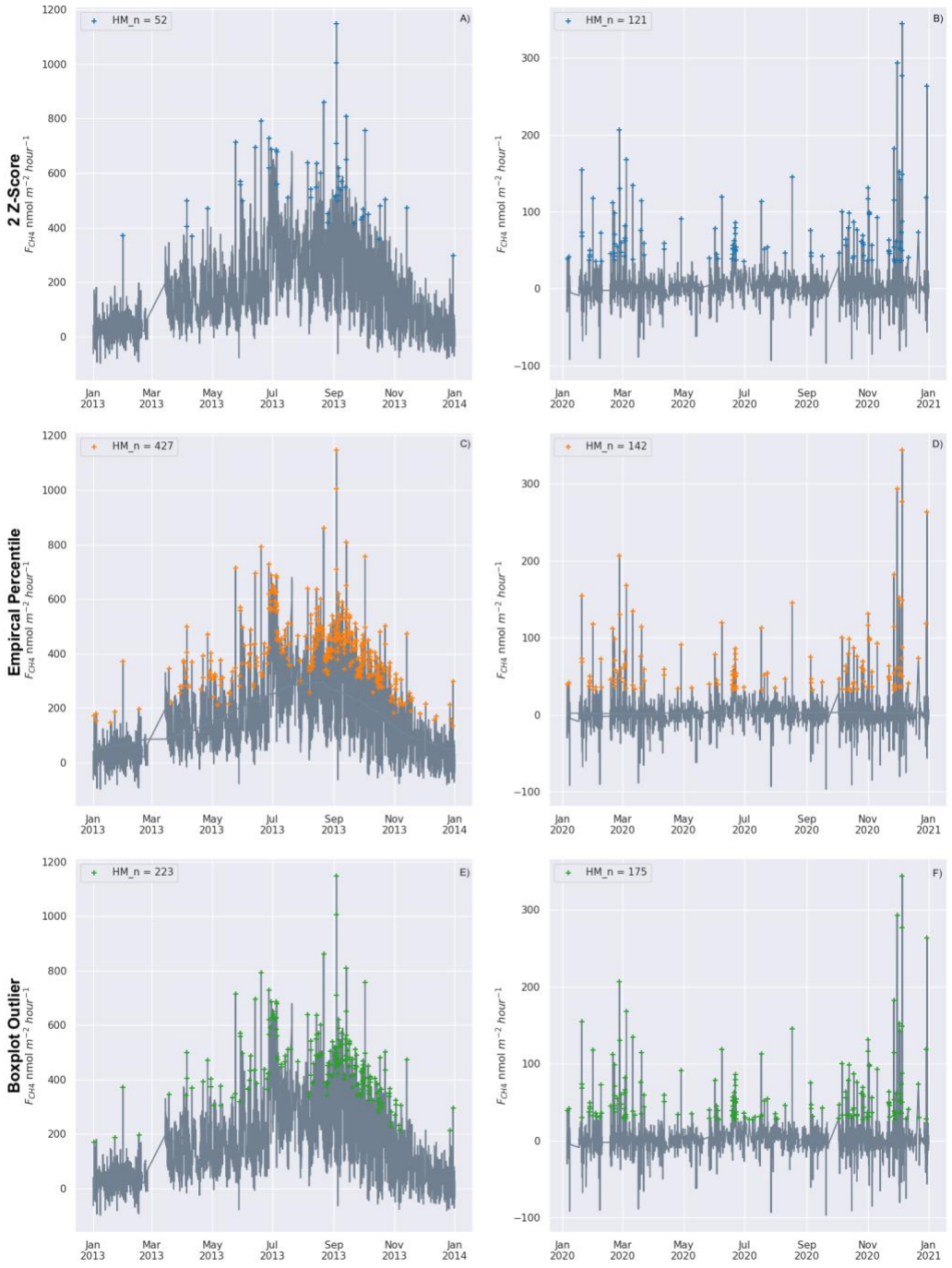


Figure 14: Comparison of Z-score, percentile, and boxplot cutoff HMs for 2013

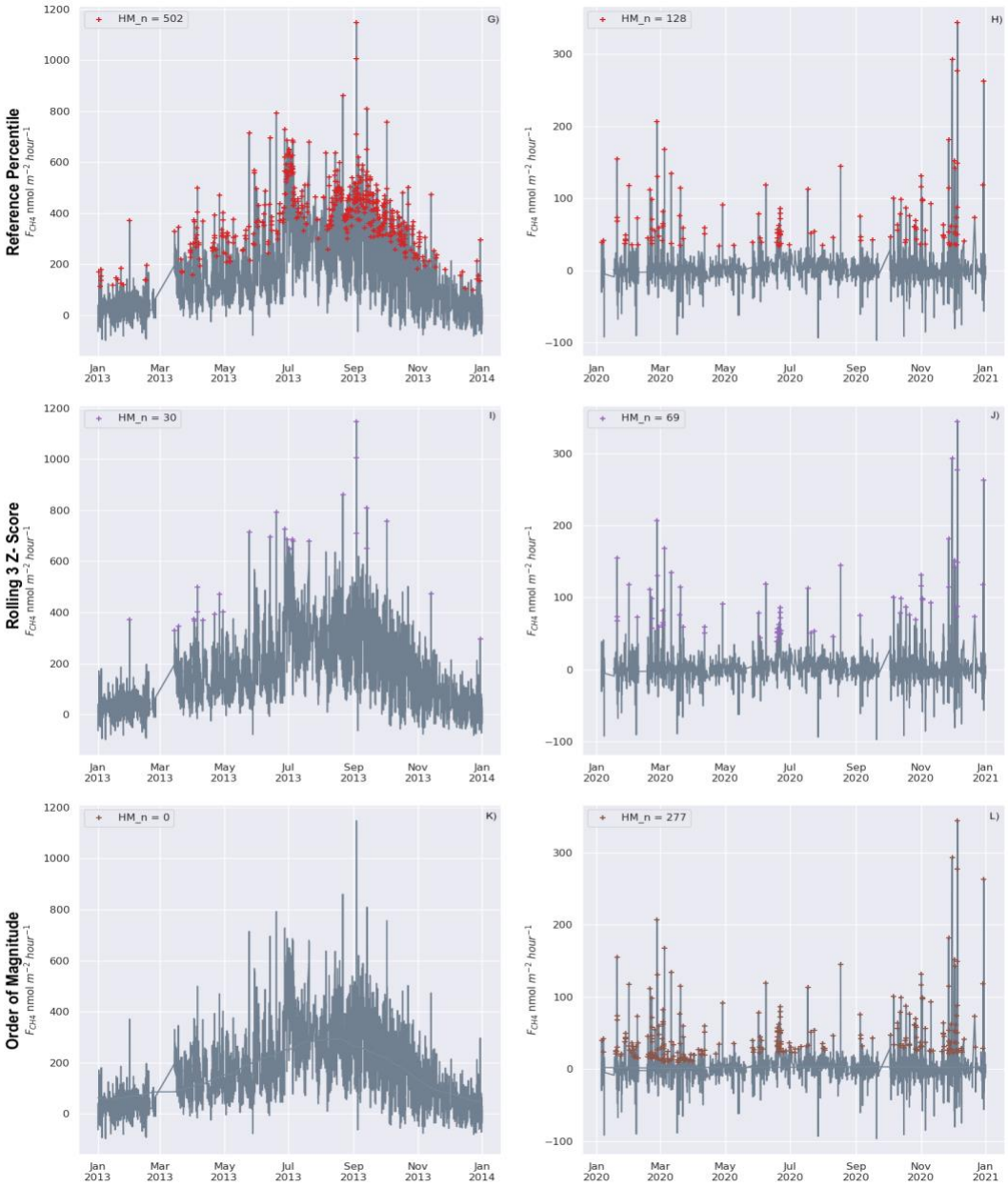


Figure 15: Comparison of reference distribution, rolling Z-score, and order of magnitude cutoff HMs for 2013 at MYB and 2020 at EDN.

4.3.1 Percentile, Boxplot Outlier, Reference Distribution Cutoffs

We first considered the adherence to the definition of HM as a way to gauge whether or not the HMs flagged by each l method are truly HM. Figures 14 and 15 show that the HM detected by the 97.5th percentile, the extreme UF, and the reference distribution 97.5th percentile cutoffs at MYB and EDN seem to over-identify HMs, and Table 3 shows the HM count for these three methods, the percentage of total measurements the HMs comprise, and the percentage of total FCH4 contributed for a single site year. While these methods flag distinct peaks in the time series as HMs, they also flag many measurements that may not be genuine hot moments and instead belong to the 'intervening' time periods of normal FCH4. These 'intervening' periods represent the times between the genuine hot moments where methane levels are within expected or typical ranges. Examining Figures 14 and 15, we can see that the 97.5th percentile cutoff inadvertently identifies some measurements as hot moments when they are just part of the normal fluctuation of methane levels. Table 3 shows that at MYB, the three methods flag between 3-6% of the data from 2013, which contribute 8-16% of total flux, indicating that the methods are capable of flagging rare events that contribute disproportionately to overall flux. However, we determined that these thresholds incorrectly capture too many normal flux measurements as HM. Compared to the visually identified HM, we present in Figure 14, which shows what measurement we deem probable HMs, we can see that

these methods strongly over-identify HM. At MYB, we determined that 38 probable HMs were present in 2013, and these methods flagged between 200 and 500 HMs. At EDN, we noted 46 probable HMs, and these three methods detected 100-270 HMs. Based on these comparisons, we determined that the 97.5th percentile, the extreme UF, and the reference distribution 97.5th percentile approaches tend to over-identify HMs. Note that the performance of these methods would likely be improved by increasing the percentile and UF cutoffs. However, to highlight the common methodologies used in the literature, we present the typical 97.5th and extreme UF cutoffs (analyses with higher HM thresholds are shown in Supplementary Materials).

The boxplot outlier method is the most common hot moment identification technique. Its benefits include its simplicity, low computation cost, and ability to handle any data distribution. However, when applied to the data from MYB and EDN, we found that even the extreme UF threshold incorrectly flagged many regular flux events as HM. This method's utility depends on the factor by which the user multiplies the IQR (Equation 4), and elevating the UF to an even higher value than the commonly used extreme UF could improve this method's performance (Molodosky et al., 2012; Johnson et al., 2010). Comparing empirical data distribution to an analogous reference distribution is conceptually straightforward and has a strong statistical grounding. However, we also found that the reference distribution percentile method tends to over-identify HM, even when we raised the HM threshold

from the default 95th percentile to the 97.5th percentile. Comparing the data to a normal distribution leads to the over-identification of HM because the 97.5th percentile of the normal distribution corresponds to a lower FCH4 value than the empirical 97.5th. The default configuration of this HM detection technique in the <hotspomoments> R package is to compare the empirical data to a normal distribution. However, perhaps when working with leptokurtic and skewed data, such as the MYB and EDN data, it would be more appropriate to use a reference distribution that is more similar to the empirical data. The empirical percentile cutoff approach also strongly over-identified HMs at MYB and EDN. The main limitation of the percentile cutoff method is that depending on the percentile threshold one chooses, it will always flag the same top percent of the data, regardless of whether all those measurements are truly distinct enough from baseline flux to be considered HMs. For example, when we set a 97.5th percentile cutoff in the raw MYB data, it returned 1826 measurements as HM, and when we set the same percentile cutoff in the de-trended MYB data, we also flagged 1826 measurements. The measurements flagged between these raw and de-trended MYB data differed (in the raw data, it was mainly the growing season fluxes), but the HM always corresponded to 2.5% of the data with the highest fluxes. Additionally, when working with big datasets, even setting a moderately conservative percentile threshold like 97.5th, will return a large number of data points because 2.5% of a dataset with >100,000 values will always yield a large number of data points that may not only include HM but rather fluxes in

the normal baseline range. In our opinion, these facets of the percentile cutoff HM detection approach make it fundamentally unreliable for HM identification.

4.3.2 Order of Magnitude Cutoffs

The order of magnitude cutoff method is distinct from the five other methods in that it does not calculate a cutoff based on the distribution of the datasets. Instead, it sets an HM cutoff as ten times higher than the seasonal mean, regardless of whether there are even measurements that high in the data. Because MYB has a strong seasonal signal and EDN does not, the order of magnitude cutoff approach had opposite results when applied to MYB and EDN. At MYB, setting the cutoff at ten times the seasonal rolling mean amplified the already strong seasonal signal, and very few HMs were flagged due to the excessively high threshold. At EDN, where the seasonal signal is between -0.9 and 3 nmol CH_4 m^{-2} hr^{-1} and HMs that can reach up to 300 nmol CH_4 m^{-2} hr^{-1}, ten times the seasonal rolling mean was not high enough to filter our regular CH4 emission events. As a result, nearly all CH4 emissions were flagged as HM. As shown in Figure 12, the order of magnitude approach yielded the most HM at END and the least (zero) HM at MYB in 2013. The results at each site fail to adhere to the HM definition for opposite reasons. The shortcomings of the order of magnitude threshold are also apparent when we compare the flagged HM to the visually detected HMs. For MYB, we manually tagged 35 HMs in 2013, and the

order of magnitude did not detect any HMs in 2013. At EDN, we manually tagged 46 probably HMs, and the order of magnitude cutoff vastly over-identified HMs with 277 in 2013 alone. The HM flagged at EDN comprised 5.5% of the tidal measurements from 2020 and contributed 122% of total measurements, indicating an extremely disproportionate influence on overall FCH4. The inconsistency in this approach is caused by the fact that when setting the order of magnitude cutoff, measurements are not considered in relation to each other, and the threshold is not calculated with the distribution of the actual data in mind.

4.3.3 Z-score Cutoffs

The Z-score hot moment identification method is a simple yet statistically robust way to identify hot moments. The classic and rolling Z-score methods provide a stricter cutoff, and the HMs flagged by these methods adhere more closely to the definition of HMs. Figures 14 and 15 illustrate that the HMs flagged by the Z-score methods are all distinct from the baseline FCH4 and do not accidentally flag many intervening, normal flux as HMs. The rolling three Z-score appears to be the approach most consistent with the visual assessment at both MYB and EDN, and each HM in 2013 is clearly anomalous from surrounding areas in the time series. Note that at EDN, since the range in FCH4 is smaller than at MYB, the Z-score methods may appear to over-identify HM. However, compared to the baseline flux that hovers

around zero, many mild HMs are in the 50-100 nmol CH_4 m^{-2} hr^{-1} range, which is still quite distinct from regular intervening flux. The classic two Z-score method flagged more HMs than the rolling Z-score approach in MYB and EDN; however, if we refer back to the entire classic Z-score time series in Figure 7, it appears that this method does not flag enough HMs, capturing only 267 HMs out of the total measurements at MYB. We determined that rolling three Z-score was more appropriate for our sites and provided a more discerning look at HMs because comparing each measurement to its' seasonal moving means more closely captures the idea that the HMs are disproportionately high compared to the intervening time periods. The rolling z-score HMs were also closest to the measurements we manually tagged as probable hot moments in 2013, which can be seen in the similar HM counts (30 flagged with Z-score cutoff and 37 visually identified at MYB; 69 flagged with Z-score cutoff and 46 identified visually at EDN) and the similarity in which measurements were flagged as HM - shown in Figures 14, 15 and 16. In the complete time series for MYB and EDN, the rolling z-score provides the most balanced approach and can flag mild and extreme HMs but does not flag intervening baseline flux measurements. At MYB, while the rolling z-score HMs only comprised 0.40% of the total measurements, they comprised 1.46% of the total FCH4. While seemingly less drastic than the disproportionality of other methods in Table 11, there is still an order of magnitude difference between the percent of total FCH4 contribution and the percent of total measurements.

We also found that the period one selects to create a rolling average timeseries dramatically affects which data points are flagged as HMs with a Z-score. Because we are using the rolling average as a conceptualization of 'baseline' normal flux from which we can compare HMs if the curve is too smooth, the rolling Z-score identification technique overidentified HMs in the growing season. If the rolling curve is too variable with many amplitude changes, we over-identify HMs throughout the year (Supplementary Materials). We identified that using a seasonal rolling average yielded a curve that was smooth enough to capture longer-term patterns in the data accurately but still variable enough to capture the fluctuations that characterize FCH4. When the seasonal rolling average was combined with a three rolling z-score threshold, we flagged the set of HM, and we feel most confident are the true HM at MYB and EDN.

5.0 Discussion and Recommendations

This paper aimed to systematically compare the most common HM identification methods highlight each method's pros and cons, identify the best-performing detection methods, and make best practice recommendations for the HSHM community. The results in the section above report the HM detected by each

method and their contribution to overall flux. Table 3 shows the significant discrepancies in how many HM were flagged and how much flux they contributed to the overall FCH4 between each detection method. These inconsistencies highlight the need for standardized HM detection best practices.

5.1 Towards an Ensemble HM Identification Approach

To achieve a more quantitative comparison of HM detection method performance, we created an HM count metric for each measurement that compiled how many HM detection methods flagged a measurement as an HM (0 meaning no methods flagged the measurement as a hot moment and 5 meaning all five methods flagged the measurement as an HM). To make this count, we excluded HM flagged by the order of magnitude cutoff, as this method's performance was weak and inconsistent. This metric dubbed the HM Count, can help conceptualize and visualize which measurements are consistently flagged as HMs. At MYB, very few hot moments were flagged by all five identification methods, and conversely, at EDN, HM flagged by all five ID methods was the most abundant. The timeseries and stems graphs in Figure 16 show the HM Count metric, and we can see that the more extreme the HM, the more likely it is to be tagged as 'hot' by all five ID methods.

The HM Counts metric also provides a case for using multiple hot moment identification methods to flag HMs at a site. Using multiple methods in conjunction to

flag HMs is akin to an ensemble classification model where different models or decision trees independently determine what class a value belongs to and vote on a final classification based on the individual model's prediction. For example, in this study, we could use the five HMID methods to assess every measurement in the FCH4 dataset and say that only if three or more of the methods flag a measurement as an HM will we conclusively assign that measurement a 'hot' label. Future work towards a robust, standardized HM identification method could explore creating an ensemble-style model that systematically applies different statistical criteria for HM detection and assigns a final 'hot' tag if the ensemble's majority vote is 'hot.'

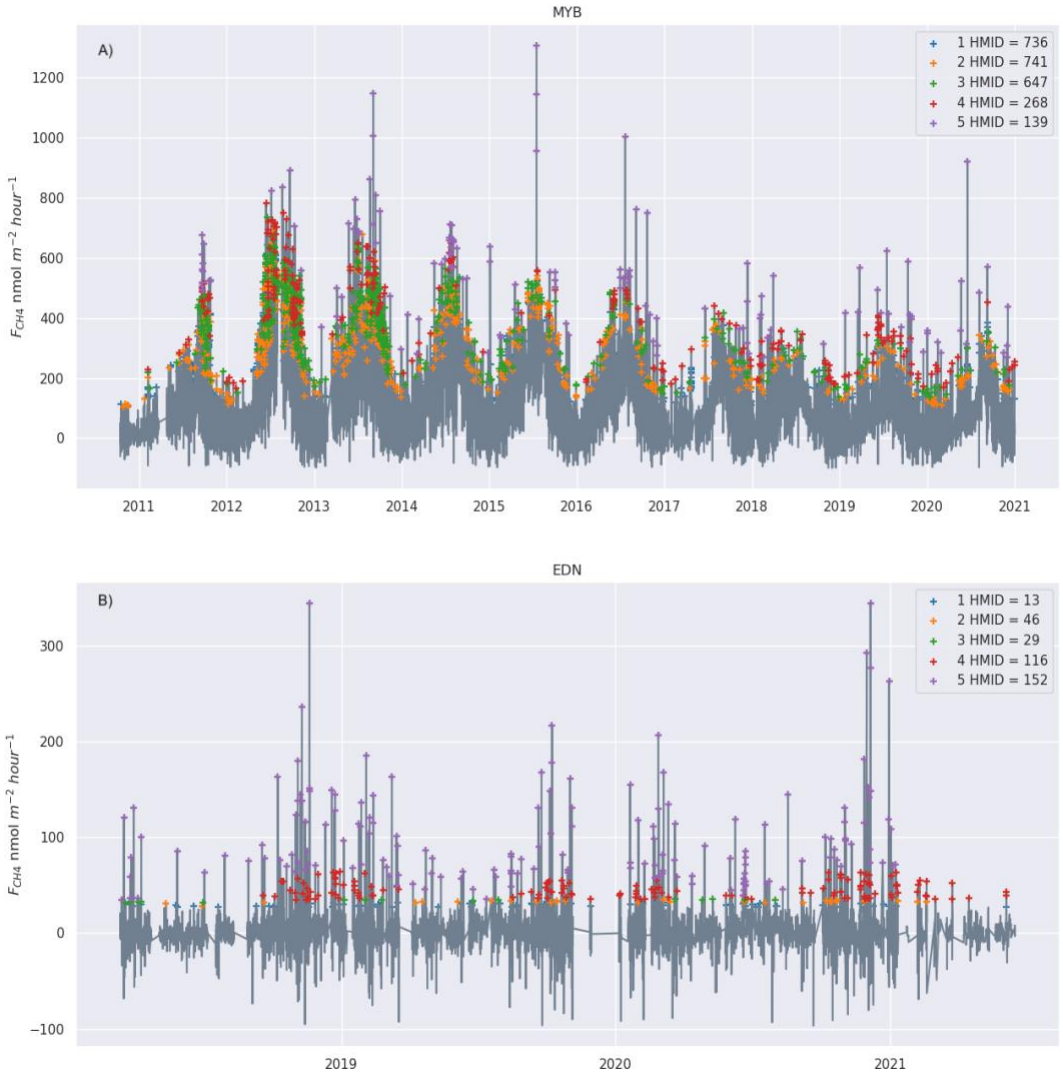


Figure 16: HM counts are displayed for MYB (A) and EDN (B). HMID stands for HM identification method, and the numbers 1 - 5 in the legend indicate how many techniques flag each measurement as an HM.

4.3 Recommended Best Practices

In this study, we reviewed several HM identification practices and assessed how

they perform when applied to FCH4 data from two wetlands in the SF Bay-Delta area. The comparative results of the study indicate that for our sites, the rolling three z-score detection technique was best suited for HM identification. We also determined that the empirical percentile and order of magnitude cutoffs are fundamentally ill-suited for detecting true HMs. The question of where to set the HM threshold is relevant to all six methodologies discussed in this study. Designating an HM cutoff is a Goldilocks problem where one wants a threshold high enough to exclude regular flux measurements but also low enough that both milder HM can still be flagged and the search window for HM is not limited. We found that the rolling Z-score approach with a three SD cutoff was our 'just right' detection threshold that allowed us to exclude baseline FCH4 variation and capture HMs that were distinctly elevated relative to baseline flux. Therefore, while we could find the best detection method for our sites, it is difficult to conclusively determine the 'best' hot moment identification technique that should be used in every study. Here, we recommend a best practice workflow that all hot moment studies should employ to ensure that hot moments are statistically sound and appropriate for the studies site/system:

1) **Perform Hot Moment Presence Tests**: We recommend starting an HM identification study by testing for the presence of HMs in a dataset with Walter et al.'s (2023) skewness and kurtosis test to determine if extreme

values dominate one's dataset. Next, one can calculate the CPI and create a Lorenz Curve and Gini Coefficient for their data to examine the degree of disproportionality in the extreme values. If these tests and visualizations indicate that extreme values contribute to an outsize effect on overall reaction rates, then one can assume they are hot moments and proceed to hot moment identification tests.

2) **Assess Data for Cyclical or Other Temporal Trends**: Once one has statistically proven the existence of hot moments in a dataset, one should move on to time series analysis and determine if any trends in the data could influence hot moment identification. Trends can complicate hot moment identification using distribution-based detection approaches like Z-score and boxplot outliers. If any trends are found in the data, one can use a trend-agnostic identification approach, such as the rolling Z-score technique, or steps should be taken to adjust or remove the trend before applying hot moment identification techniques that rely on overall data distribution.

3) **Apply Hot Moment Identification Techniques:** After exploratory data analysis and removing cyclical signals, if applicable, one can start applying hot moment identification techniques to their data. We recommend choosing one of the statistically grounded methods tested in this study instead of qualitatively flagging hot moments. Using one technique over another might

be more appropriate depending on site dynamics. However, we recommend against using the percentile and order of magnitude cutoffs as they exhibited weak and inconsistent performance in this study. Instead, we suggest that researchers use the boxplot outlier, reference distribution percentile, and rolling and classic z-score approaches and experiment with the HM threshold in each approach to find the optimal technique and HM cutoff for one's data. We argue it is best practice to try various hot moment identification methods to test for agreement between which measurements are flagged as hot moments and build confidence in your 'hot' assignments.

Chapter 2: Predicting Hot Moments of FCH4

Introduction

1.1 Coastal Wetlands and the Carbon Cycle

Coastal wetlands play a critical role in the global carbon (C) cycle and are incredibly efficient at capturing and sequestering C (Sutton-Grier et al., 2014), storing 1 to 3 orders of magnitude more C in their sediments than freshwater wetlands and forests (McLeod et al., 2011; Rosentreter et al., 2021). Wetlands can influence Earth's radiative forcing by removing atmospheric carbon dioxide and storing C in

the soil and vegetation. High plant productivity and low organic matter decomposition rates in these ecosystems make C accumulation in wetland soils very high. Because of their ability to sequester large amounts of C and promote radiative cooling, wetlands restoration projects have been proposed as a management strategy to draw down C and mitigate the effects of anthropogenic climate change (Callaway et al., 2012).

Wetland restoration projects have become increasingly popular in the San Francisco Bay-San Joaquin River Delta (Bay-Delta) region in Northern California, USA, with more than 30,000 acres in the region in various stages of restoration (California Natural Resource Agency, 2018). In recent years, the state of California and other stakeholders have undertaken ambitious plans to restore more wetlands in the Delta (EcoAtlas, 2013). Farmland conversion to wetlands is particularly attractive in this region, where coastal wetlands can store lots of C compared to drained farmland, emitting around 44 Gt of C annually (Hemes et al., 2019). In the late 1850s, the Bay Delta's dominant land cover was tidally influenced brackish and freshwater wetlands, covering about 1,300 $km^{2,}$ and about 87% of the region's total area was made of carbon-rich peat soil. However, over the last 150 years, 85% of the wetlands in the Delta have been drained or diked to support agriculture (Coastal Conservancy of California, 2010). This dramatic land use and land cover changes have led to a robust agricultural economy in the area, but also widespread soil subsidence and

increases in GHG emissions from drained peatland (Hatala et al., 2012; Ingebritsen et al., 2009). Restoration projects in the Bay Delta typically involve flooding farmland, increasing water retention, and excavating or constructing berms (South Bay Salt Pond Restoration Project, 2007). In some cases, wetland managers plant native, historic vegetation in the newly flooded or filled wetlands to restore the farmland as close to each site-specific historical ecology as possible.

1.2 The Methane Compromise

Wetland restoration is an attractive and promising strategy for mitigating climate change; however, the same anaerobic soil conditions that allow coastal wetlands to store large amounts of C efficiently also make coastal wetlands ideal for methanogenesis (Neubauer & Megonigal, 2015; Poffenbarger et al., 2011). In fact, CH4 emissions from wetlands have been identified as the leading natural source of this highly potent greenhouse gas (GHG) in the atmosphere (Saunois et al., 2017). Therefore, wetland restoration is often accompanied by a "methane compromise" as these systems can produce and emit large amounts of CH4 (Hatala et al., 2012; Hemes et al., 2018; 2019). One way to conceptualize the balance between C storage and CH4 emissions in restored wetlands is by measuring or predicting the FCH4 for a wetland system. While wetlands commonly exhibit negative net radiative forcing on geologic timescales (Frokling & Roulet, 2007; Frokling et al., 2006; Mitsch et al.,

2013), on shorter time scales more relevant to human activities, CH4 emissions can significantly impact radiative forcing, contribute to atmospheric warming, and potentially reduce the efficacy of these systems as C sequestration climate mitigation measures (Gedney et al., 2004; Holm et al., 2016). It would, therefore, be useful to understand better the factors or conditions that result in enhanced FCH4 and design restoration projects that strive to reduce the occurrence of such conditions. Indeed, the efficacy of wetland restoration for C sequestration is under investigation in many wetlands across the US (Hemes et al., 2019; Griscom et al., 2017).

While there are thousands of acres of restored wetlands and many more planned in the Bay-Delta, studies that have examined the question of CH4 budget in restored wetland systems in the Bay-Delta have mixed results regarding the exact ways restoration alerts the CH4 emission patterns (Anderson et al., 2016). For example, Knox et al. (2015) found that newly restored wetlands in the San Francisco Bay were net sinks of CO2, sequestering up to 397 g C m-2 yr-1, but also a significant source of CH4, emitting 39-53 C m-2 yr-1. Using a simple model of radiative forcing and atmospheric lifetimes, Hemes et al. (2019) showed that restored peat soil-managed wetlands do not become net sinks of GHG until a century post-restoration. However, other studies report low CH4 emissions in earlier inundation stages during wetland establishment (Hatala et al., 2012; Hahn-Schofl et al., 2011). The striking variability of emissions throughout the restoration timeline at any

specific wetland and between different restored wetlands is significant and may complicate including wetland restoration in carbon offset programs (Cook, 2016).

1.3 Hot Spots and HMs of FCH4

CH4 production and emission in the Bay-Delta and beyond are controlled by diverse biogeochemical conditions, many of which are affected by management at each wetland site. Such conditions include vegetation type, water table elevation, soil water content, air temperature, soil temperature, oxygen saturation, salinity, nutrient availability, and soil accumulation rate. CH4 emissions typically follow two pathways from the wetland to the atmosphere: plant-mediated transport and diffusion or ebullition from air/soil and air/water interfaces in a wetland. Coastal wetlands are often a patchwork of vegetation, mudflats, and open water and thus exhibit substantial biogeochemical heterogeneity and can transport CH4 through multiple pathways at once (Hunt et al., 1997; Alongi, 2020). These CH4 transport pathways operate on different time scales and respond to external forcing from biogeochemical conditions at different rates (Turner et al., 2020).

The complex suite of controls on CH4 cycling in coastal wetlands leads to uneven emissions in a wetland across space and time and elevated CH4 flux (FCH4) at certain times and from specific locations. We can contextualize these extreme CH4

emission events and locations using as hot spots and hot moments (McClain et al., 2003). HSHM of CH4 emission have been documented in numerous wetlands around the globe (Waldo et al., 2020; Obregon et al., 2023; Tupek et al., 2015) and in the Bay-Delta region (Rey-Sanchez et al., 2022; Anthony et al., 2023). While a large body of work has quantified wetland CH4 flux and identified drivers of emissions at the daily and annual scale (Liu et al., 2019; Jeffery et al., 2019), the way these drivers scale spatially and temporally is poorly understood (Saunois et al., 2017). Moving across scales to examine short-term CH4 emission events induced by accelerated biogeochemical process rates and honing into specific zones in coastal wetlands that experience heightened flux will allow for a deeper understanding of coastal wetland FCH4, more accurate estimates of flux and modeling future C cycling dynamics in these systems and could also inform restoration projects.

1.4 Modeling and Predicting Extreme CH4 Flux

The variability in CH4 budgets is not unique to the Delta, and quantifying global CH4 emissions from coastal wetlands has been characterized by significant uncertainties. Estimates of CH4 emissions from coastal wetlands were initially thought to range from 1-3 Tg CH4 yr^{-1} (Bange et al., 1994; Upstill-Goddard et al., 2000; Middelburg et al., 2002). The 2017 Global CH4 Budget Report (Saunois et al., 2017) did not include partitioned CH4 emissions from coastal wetlands. However, the

most recent iteration of the report synthesized global estimates and determined that coastal wetlands emit between 4 – 5 Tg CH4 yr^{-1} (Suanois et al., 2020). The high uncertainty present in coastal wetland emission estimation has been attributed to conflicting designations of coastal ecosystem types, which can lead to overcounting and undercounting of CH4 emissions, uncertain estimates/measurements of coastal wetland surface area, and poorly quantified CH4 emissions rates from the various coastal ecosystem types (Suanois et al., 2020).

Additionally, the spatiotemporal heterogeneity that characterizes coastal wetlands makes it difficult to generalize and upscale measurements for other wetlands. HSHMs of FCH4 have been notably absent when estimating and modeling the C balance of coastal wetlands. Because stochastic HSHM often dominates coastal wetland CH4 dynamics, the notable absence of HSHM likely contributes to the uncertainty in coastal wetland CH4 budget estimates in the Delta and globally. Utilizing the HSHM framework and identifying and quantifying HSHM in an ecosystem ensures that the spatiotemporal dimensions of extreme values that often have a disproportionate influence on overall reaction rates and ecosystem functioning and are critical for accurate modeling and upscaling approaches are not lost (Walter et al., 2020; Bernhardt et al., 2017).

Machine-learning algorithms have been used to upscale various EC data,

including CH4 flux, to estimate CH4 freshwater wetland emissions (Peltola et al., 2019; Knox et al., 2019) and identify the dominant drivers of FCH4 in a mangrove forest (Liu et al., 2019). Data-driven models are a versatile tool that has been used to combine EC tower measurements with remote sensing observation and climate models to upscale fluxes (McNicol et al., 2023; Huang et al., 2021; Tramontana et al., 2015). However, we have yet to encounter any data-driven models specifically focusing on predicting and upscaling HM of FCH4. HSHMs are generally underrepresented in predictive models that either omit or inadequately capture the interactions resulting in HSHM (Walter et al., 2023). Because HMs of FCH4 can occur on an hourly scale in coastal wetlands (Pearsall et al., 2024), monitoring HM of FCH4 requires high temporal resolution tools such as EC that can capture these rare occurrences. However, EC towers are expensive to install and logistically unfeasible in every wetland. In coastal wetlands, most CH4 monitoring is instead done with chamber measurements (Al-Haj & Fulweiler, 2020). Chamber measurements are typically taken at much lower frequency, making the temporal resolution of chamber measurements too low to capture the frequency of rare HMs accurately. Another benefit of EC towers is that they collect a suite of standard environmental variables, along with energy and GHG fluxes, and we can link/model the relationships of these variables to HMs of FCH4. Other, more common ecological monitoring, such as weather stations and remote sensing, also collect this suite of environmental parameters. Therefore, upscaling EC data is especially useful for HM of FCH4 in

coastal wetlands, and by extrapolating from widely available environmental parameters, an upscaled model enables the prediction of HM in diverse wetland settings, even in the absence of dedicated EC towers.

Data-driven approaches excel at discovering and extracting patterns from data without prior knowledge of the system and can provide valuable insights into complicated natural systems (Montáns et al., 2019). The RF methodology has been employed to forecast other extreme events analogous to HSHM, such as marine heat waves, flooding, and extreme precipitation (Giamlacki et al., 2021; Schumacher et al., 2021; Herman & Schumacher, 2018). The RF methodology is particularly advantageous for wetland FCH4 inquiry as the complicated relationships between biogeochemical parameters were learned rather than assumed. Our objective in this study was to assess the potential for upscaling and predictability of HM using a RF model. We built several differently parameterized RF classifiers that predicted HM absence or presence on the hourly scale and identified the best-performing model. We also determined the most important (measurable by less costly and more prevalent methods) predictors of HMs, which shed light on the potential for upscaling our model to wetlands without EC towers. We utilized open-access data from nine wetlands in the Bay Delta area to train the RFs and assess their performance and feature importance to determine their HM predictive utility.

Data and Methods

We compiled data from nine wetland sites from the Bay-Delta region that are all part of the AmeriFlux network as training data for the RF model. The nine sites of interest in this study compose a network of varied hydrological and ecological wetlands, with four of the five sites being tidally influenced (Map 1). The sites in this study are also a mix of restored and undisturbed wetlands and marshes. Each site has a unique hydrological history and specialized management plan that influences each wetland's biogeochemical conditions and drivers of FCH4. Because of each wetland's varied biogeochemical and structural characteristics, this network of sites has a wide range of GHG emissions. Furthermore, since most of the wetlands in this study are restored, their emission and source/sink change dramatically over time as restoration projects progress. Differences in bathymetry and impounding techniques lead to differences in each site's open water vs. vegetation makeup - which also changes as restoration progresses and vegetation re-establishes itself at each restored site. Additionally, salinity, water level, and tidal influence can impact GHG emissions, explaining some of the variability in GHG emissions among wetlands in the Bay-Delta regions. The network consists of one historic wetland in Suisun Ranch Reserve and eight restored and managed wetlands: three restored wetlands on Twitchell Island, a restored wetland near Mayberry Slough on Sherman Island, a restored marsh

and salt pond in Eden Landing, a restored peatland pasture on Sherman Island, a recently flooded tidal wetland at Hill Sough, and an impounded and flooded tidal wetland at Dutch Slough.

Mayberry Wetland (MYB), a 121-hectare restored area on Sherman Island, was transformed from a livestock pasture dominated by pickleweed into a wetland in 2010 through a restoration project managed by the Department of Water Resources. This site's landscape is 50% open water and 50% vegetation, with variable water depths and vegetation patterns reflecting the heterogeneous bathymetry. Managed by the California DWR, water from the nearby river is piped in during dry summers to maintain water levels (Arias-Ortiz et al., 2022).

Sherman Island Wetland (SNE), spanning 263 hectares on the southwestern side of Sherman Island, underwent restoration from a degraded peatland pasture in 2016. Despite being inundated in November 2016, widespread vegetation establishment was slow, with tulle and cattail only covering the site by 2020 (Hemes et al., 2019). SNE's water levels are managed through a pump system due to its lack of natural hydrologic connectivity.

Twitchell Island's East End Wetland (TW4), spanning 323 hectares, was restored from a cornfield in continuous agricultural use since the 1850s, with

construction starting in December 2013. The site now features a combination of open water and emergent vegetation, primarily tulles and cattails. It is managed by a water level regime utilizing pumps to maintain ideal inundation levels (Valach et al., 2020; Hemes et al., 2018).

The West Pond Wetland, a 3-hectare restored site on Twitchell Island managed by the California Department of Water Resources (CDWR), was restored from a degraded cornfield in 1997. Flooded to a depth of 25cm during restoration, pumps, inlets, and outlets still regulate the hydrology at this site. With shallow and uniform bathymetry, dense vegetation, primarily cattails and tulles, quickly established after restoration, covering 96% of the wetland area by 2012 (AmeriFlux).

The 6.5-acre East Pond Wetland on Twitchell Island (TW5), restored from drained agricultural fields in 1997, is jointly managed by the CDWR and the U.S. Geological Survey (USGS). Initially flooded to approximately 55 cm, it is still hydrologically managed using pumps and inlets. After restoration, vegetation, mainly tulles and cattails, gradually established across the wetland. However, a disturbance occurred in 2013 when vegetation was harvested and relocated to facilitate the restoration of the East End wetland.

Rush Ranch (SRR) is a 425-acre undisturbed wetland in Suisun Marsh in Suisun Bay, managed by CDWR as part of the San Francisco Bay National Estuarine Research Reserve. It stands out as the largest continuous brackish marsh complex in

California and features pickleweed, cordgrass, and tulles as the dominant vegetation. The site experiences tidal influence and has a complex hydrology (SF Final Management Plan, 2020).

Mount Eden Landing Creek Marsh (EDN) is a 75-hectare tidal marsh in the Eden Landing Ecological Reserve, managed by the CDFW. Restored from salt ponds in 2008, it features mudflats and vegetated areas dominated by pickleweed and cordgrass. Despite lower vegetation, significant spatial and temporal variability in CH4 emissions is observed, with peaks exceeding 200 nmol CH4 m-2 hour-1, contrasting with the seasonality of emissions seen in Mayberry Wetland (Shahan et al., 2022; Arias-Ortiz et al., 2023).

Dutch Slough Marsh (DMG), a 0.41-hectare restored tidal marsh in Suisun Bay, was previously used for agriculture and slated for urban development. In 2018, the CDWR leveled the slope, planted vegetation, and breached levees to reconnect it to the Delta. Data collection for this project began after flooding and tidal reconnection (AmeriFlux)

Hill Slough Marsh (HSM), a 0.40-hectare restored tidal marsh in Suisun Bay's northern Suisun Marsh, was previously eight diked salt ponds managed for waterfowl habitat. Restoration, managed by Ducks Unlimited Inc. and the CDWR, began in 2017, involving levee modifications and flooding in 2021 to reintroduce tidal action. Data used in the study was collected after the restoration's completion, and tidal

connectivity was restored at this site (AmeriFlux)

Table 4: Bay-Delta Site Summary

	Site ID	Location	Data Years	Salinity (PSU)	HM Count	HM% of total Measurements	HM% of total FCH4
Non-Tidal	US-MYB	Sherman Island	2010-2021	1.0-7	1691	2.32%	6.41%
	US-SNE	Sherman Island,	2016-2020	1-7	840	2.84%	9.3%
	US-TW1	Twitchell Island	2011-2020	0.1 - 0.3	931	2.46%	7.43%
	US-TW4	Twitchell Island	2013-2020	0.1 - 0.3	1243	2.43%	6.83%
	US-TW5	Twitchell Island	2018-2020	0.1 - 0.3	212	2.47%	4.66%
Tidal	US-SRR	Suisun Bay,	2014-2017	3-13	552	3.0%	20.1%
	US-EDN	South San Francisco Bay	2018-2020	30-35	326	2.38%	92.86%
	US-HSM	Suisun Bay,	2021-2023	3-13	290	3.26%	39.92%
	US-DMG	Oakley, California	2021-2022	0.1-0.6	303	3.1%	12.87%

2.1 Data Sources and Preparation

The CH4 flux data utilized in this study was collected using the eddy covariance (EC) flux method. EC is a micro-meteorological method that directly observes gas, energy, and momentum exchanges between ecosystems and the atmosphere using a flux tower monitoring system deployed to field sites of interest (Baldocchi, 2003). EC has few theoretical assumptions and an extensive scope of application, and towers have been deployed in a wide range of ecosystems. The data

used in this study was accessed through the AmeriFlux open-access database, a network of 110 active PI-managed sites measuring ecosystem CO2, CH4, water, and energy fluxes in North, Central, and South America. Each ecosystem-level site in the AmeriFlux network acquires continuous EC measurements and metadata describing biological, ecological, and management conditions at each site, such as climate zone and dominant vegetation type. All data published in the AmeriFlux network undergoes rigorous processing, which includes data quality control and transforming original measurements from individual sensors to ecologically or micro-meteorologically significant quantities (Knox et al., 2019). Since the AmeriFlux network data has already undergone quality control and assurance from each site's PI before uploading it to the database, pre-processing of the data was minimal in this study. All data was downloaded at the half-hour resolution, and to reduce noise, we aggregated half-hourly measurements into hourly measurements.

To make the model relevant for upscaling across sites without EC towers, we only used HM predictors that could be collected by alternative means, such as weather stations or remote sensing. While we omitted FCH4 in our predictors, we used the FCH4 measurements to flag HMs, and all the data used was collected contemporaneously with FCH4 measurements. Therefore, with the HMs as our target variable, RFs can learn the relationships between elevated FCH4 and the other environmental parameters. We also limited the predictors to only those available at

each site, and the list of variables used in the modeling is presented below. The HM and tidal flags were assigned to each measurement at the nine sites, and they denote whether a measurement is an HM or a non-HM and if the measurement is from a tidal or non-tidal site. The HM flag was our target variable that the RF tried to predict. We included a tidal/non-tidal flag in the dataset to assess if tidal action at a site influenced HM prediction. For instance, if we ran the RF and found that the tidal flag was an essential feature in the model, we could infer that tidal action plays a prominent role in HM occurrence and that it might be more appropriate to have separate tidal and non-tidal RF models. We also transformed the wind direction parameter using a sine function because wind direction is a circular variable where extreme values have similar meanings. For example, winds coming from 360º and 1º in the same general direction, even though their numerical values are distinct.

Table 5: EC Variables Included in RF as Predictors

Variable	Name	Units
WS	Wind Speed	$m\ s^{-1}$
WD	Wind Direction (Transformed)	Decimal degrees
WTD	Water Table Depth	m

TA	Air Temperature	deg C
PA	Air Pressure	kPa
RH	Relative Humidity	%
TS_AVG	Soil Temperature	deg C
VPD	Vapor Pressure Deficit	hPa
PPFD_IN	Photosynthetic Photon Density, incoming	µmolPhoton m^{-2} s^{-1}
LE	Latent Heat Flux	W m^{-2}
H	Sensible Heat Flux	W m^{-2}
NETRAD	Net Radiation	W m^{-2}
H2O	Water Vapor in mole fraction of wet air	mmolH2O mol^{-1}
Tidal Flag	Flag denoting if a measurement is from a tidal or a non-tidal site	True (Tidal)/ False (Non-Tidal)

2.2 Identifying HMs of FCH4

We identified HMs of FCH4 at the nine sites in this study using a rolling average and rolling z-score methodology described in Chapter 1. To that end, we established a baseline flux by employing a moving window technique to smooth the hourly FCH4 time series into a monthly moving average. Subsequently, another

moving window was utilized to compute the moving Z-score for each FCH4 measurement relative to the seasonal rolling average for each measurement. We identified every FCH4 measurement with a moving Z-score greater than three as a HM. This approach was informed by the standard Z-score cutoff approach for identifying HMs (Kannenburg et al., 2021), moving Z-scores used to identify anomalously high stock prices in financial analysis (Velazques, 2019), and moving window outlier detection techniques often used in ecological studies like the Hampel filter (Hampel, 1971). We quantified the impact these HMs have on the overall FCH4 at each site by comparing the contribution of total HMs attributed to FCH4 to total FCH4 at each site (% contribution) (see Pearsall et al., 2024 for details).

At each site, we attached a 'hot or not' flag to each measurement, stored as a Boolean variable to denote the class to which each measurement belongs. We then compiled the data from each site into one dataset that will be used to train and test the RFs. This full Delta composite dataset did not include any site identification information, and the dataset was randomly sampled to ensure site-specific artifacts did not make their way into the final HM prediction RF. The final dataset included 4,416 HMs and 177,256 non-HMs for 181,672 measurements. HMs comprised 2.28% of the HM dataset. We also created separate tidal and non-tidal versions of the HM dataset so that we could evaluate the role tidal action plays in driving HMs. The tidal

dataset included US-EDN, US-SRR, US-DMG, and US-HSM, and the non-tidal site included US-MYB, US-TW1, US-TW4, US-TW5, and US-SNE. There were 4,856 HMs in the non-tidal dataset and 1,471 HMs in the tidal dataset.

2.3 RF Training and Optimization

A RF (Breiman, 2001) is a non-parametric, supervised learning technique that uses an ensemble of unpruned classification or regression trees to make predictions (Oshiro et al., 2012) based on predictor variables fed into the model. An RF makes predictions by aggregating the predictions of each tree in the ensemble, and in an RF classifier, a majority vote among the trees determines the final classification. In an RF, each tree is grown from a bootstrap sample, selected with replacement from the training data (Chen et al., 2004), in a process known as bagging that allows the forest to be made of unique, uncorrelated trees. At each decision tree node, a certain number of predictors are randomly selected, and the algorithm evaluates all potential thresholds for each chosen predictor and selects the predictor-threshold combination that yields the most effective split of the data (best split). In each tree, the resulting groups are split to maximize homogeneity within groups and the differences between groups (Rigatti et al., 2017). The randomization introduced during the bagging and tree seeding steps ensures trees are protected from overfitting and from individual errors in the trees. RF produces more accurate predictions than any singular decision

tree classifiers like CART and C4.5 (Prasad et al., 2006; Chen et al., 2004). Training data alone does not dictate RF performance, and model performance can be influenced by model specifications, called hyper-parameters, that determine the structure of individual decision trees and the size and randomness of the overall forest (Probst & Boulesteix, 2018). Hyperparameters are not learned from the data but are set by the user before model training. Hyperparameters can be experimentally tuned to optimize the performance of the RF model. The most effective approach to identifying the optimal settings is to try numerous combinations and assess the performance of each mode.

To build the RFs in this study, we randomly split the HM dataset into a training and a testing subset, saving 20% of the data for testing. Testing the model on unseen data reduces the overfitting change and estimates how well the model generalizes to new, unseen instances. Test segments of data help assess whether the model has learned meaningful patterns from the training data or simply memorizes the training examples (overfitting). We tested different combinations of hyperparameters using a random grid search K-fold cross-validation algorithm. We searched for the optimal hyperparameters with this algorithm: 1) maximum depth of the tree, 2) minimum number of samples required to be at a leaf node, 3) The minimum number of samples required to split a leaf, 4) number of features to consider when looking for the best split, and 5) number of trees in the forest 6)

whether or not bootstrapping should be used to build the forest. With the random grid search, we first defined a grid of six hyperparameter values we wanted to test. Then, we randomly selected a predefined number of hyperparameter combinations to try instead of exhaustively searching through all possible combinations in the grid. For each selected combination of hyperparameters, we split the dataset into K folds, trained the model on the k-1 fold, and tested the model on the remaining fold. This process is repeated K times and yields K difference models, and the average performance metric across all folds is calculated for each hyperparameter combination. From the results of the K-fold search, we select the combination of hyperparameters that yielded the highest average performance metric. Our random search K-fold cross-validation evaluation was executed with three folds, and 300 different fits of the RF were tested. We then used these optimal hyperparameters as determined by the cross-validation to build an RF trained with the balanced and unbalanced HM training dataset. To contextualize the results of the RF in this study against less sophisticated models, we trained a logistic regression, a C4.5 classifier, and a CART classifier. We compared the performance of our RF to each of these models.

Since decision trees are susceptible to the data they are trained on, small changes in the training set can significantly change tree structures. This capability is exploited by the RF method, allows for the growth of wildly different trees, and

protects against overfitting and errors (Prasad et al., 2006). However, suppose there is a class imbalance in the training data where the ratio between the majority and minority classes is highly skewed toward the majority class. In that case, RF models can become biased against identifying minority class instances. RF is especially sensitive to imbalanced datasets because decision trees use a 'divide and conquer' approach that partitions training data into smaller and smaller pieces. Majority class prediction bias causes a problem when identifying rare patterns in a dataset because there is less and less data in each tree node from which to identify rare patterns (He and Ma, 2013). HMs are inherently rare in any ecosystem, and in our wetland HM dataset, HMs comprise less than 3% of the total dataset. The class imbalance between HMs and non-HMs severely skewed RF trained with imbalanced data, and we had to apply several techniques to balance our data and account for the class imbalance to develop an RF that accurately predicted both HMs and non-HMs.

2.4 Dealing with Imbalanced Training Data

Imbalanced data, datasets where one class significantly outnumbers another, has attracted significant attention in machine learning as many real-world datasets and ML applications are significantly skewed. Famous examples in the field include breast cancer prediction datasets (Zhang et al., 2019), where negative diagnoses significantly outnumber positive diagnoses, and financial fraud data (Panigrahi &

Borah, 2019), where legitimate transactions outnumber fraudulent transactions. Most ML practitioners agree that a dataset where the most common class is more than twice as common as the rarest class is unbalanced. There can be varying degrees of imbalance - a dataset with a class instance ratio of 10:1 would be marginally unbalanced. In contrast, a dataset with a class instance ratio of 1000:1 would be severely unbalanced (He & Ma, 2013).

The problem with imbalanced data is that using highly skewed classes in data significantly compromises the performance of most standard ML algorithms because the sheer volume of the majority class pattern obscures the rarer class pattern. This imbalance makes it fundamentally more difficult for an algorithm to identify rare patterns than common patterns, and as the class imbalance in a training dataset increases, so does the model's error rate (He & Ma, 2013). When decision trees, such as RFs, are trained with imbalanced data, they can become biased against accurately identifying rare class cases because trees are built top-down (Chen et al., 2004). If one branch has little or no training examples of the minority class, the tree will have no evidence to base a classification when it encounters a minority class example in the test data subset (He and Ma, 2013; Lin et al., 2017). Therefore, when trained on imbalanced data, most decision tree learners will predict the majority class more accurately and bias the results against rarer classes. As the dataset becomes larger, the problems associated with class imbalance become more significant, and with the HM

dataset of 100,000+ instances, addressing the imbalance is imperative for improved algorithm performance. To improve the RF's ability to predict HMs correctly, we need to expose it to more HMs during model training so that it can effectively learn the decision boundary. Since each method changes the distribution of the dataset, how you balance the data will likely influence how the model learns to identify HM and feature importance in the model. There are three main approaches for dealing with imbalanced training data: 1) sampling methods solutions, 2) cost-sensitive models, and 3) algorithm-level solutions. We tested each approach in this study and determined that the best method for balancing the data was randomly undersampling non-hot moments in the training dataset so that there was an even number of HM and non-HM. Full results from our imbalanced data comparison are presented in Supplementary Materials.

Randomly balancing the training data for RF has been used in many medical and statistical studies (Zhang et al., 2019) as well as in environmental studies such as Giamalaki et al. (2021). Class balance with random undersampling is achieved by randomly selecting majority class instances and removing them from the dataset until the desired ratio of minority and majority classes is reached. Random undersampling is a naive resampling method because it does not assume anything about the data when removing majority class samples (He and Ma 2013). The computational simplicity of random undersampling makes it easy to implement and fast to execute,

which is very beneficial when working with large datasets like our FCH4 dataset. We performed the random undersampling using the <Imblearn> package in Python. After randomly undersampling the dataset, the new balanced training dataset had HM 4,416 and 4,416 non-HM.

2.5 RF Performance Measures

We used a suite of machine learning performance metrics to compare the prediction skills of the various RFs we built in this study. Some typical RF performance metrics, such as overall classification accuracy and the receiver operator curve (ROC), can be optimistic and overstate a model's performance when there is a severe class imbalance. These metrics can be skewed by high majority class prediction accuracy because low overall error can be achieved even by a no-skill RF model that can only predict majority class instances. We still report the accuracy and ROC Curve metrics for the RFs in this study but also present the balanced accuracy and precision-recall curve (PRC), which are more suitable for imbalanced datasets. We also use confusion matrices, precision, recall, and the F1-Score to compare model performance.

The confusion matrix reports the actual class labels vs. the predicted ones for MH and non-HMs. This matrix summarizes the True Negatives (TN) correctly labeled non-HMs, True Positives (TP) correctly labeled HMs, False Negatives (FN) incorrectly labeled non-HMs, and False Positives (FP) (incorrectly labeled HMs. We

also report the TN, TR, FP, and FN rates on the confusion matrix.

$$TPR = \frac{TP}{Actual\ Positives} = \frac{TP}{TP + FN}$$

$$FNR = \frac{FP}{Actual\ Positives} = \frac{TN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negatives} = \frac{TN}{TP + FP}$$

$$FPR = \frac{FP}{Actual\ Negatives} = \frac{FP}{TN + FP}$$

Accuracy is the ratio of the total number of correct predictions and the total number of predictions. Because accuracy includes the number of correct predictions and places more weight on common classes, we found that this metric severely exaggerated model performance with our imbalanced HM dataset.

$$Acc = \frac{correct\ predictions}{all\ predictions} = \frac{TP + TN}{TP + FP + TN + FN}$$

Balanced accuracy calculates a model's overall accuracy by computing the average percentage of correct positive class predictions and the percentage of correct negative class predictions. This metric assesses the performance of models trained with imbalanced datasets. Balanced Accuracy can also be calculated as the arithmetic means of sensitivity (TPR) and specificity (TNR -1).

$$Balanced\ Acc\ = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) = \frac{sensitivity + specificity}{2}$$

The ROC Curve is a probability curve that plots the TPR (sensitivity) against FPR (1 -sensitivity) and shows the performance of a classifier across different classification thresholds (Nahm, 2022) and the trade-off between sensitivity and specificity (Saito & Rehmsmeier, 2015). The area under the ROC Curve (AUC) measures the ability of a binary classifier to distinguish between classes and is used as a summary of the ROC curve (Hoo et al., 2017). The AUC ranges from 0 to 1, where a perfect classifier would have a score of 0, and a no-skill classifier that cannot distinguish between classes would have an AUC = 0.5.

ROC Curves and AUC can be too optimistic when dealing with imbalanced datasets and overstate a model's ability to discern between classes. An alternative to the ROC Curve favored for imbalanced datasets is the precision-recall curve, which

focuses on the performance of the minority class (Brownlee, 2021; Saito & Rehmsmeier, 2015). Precision quantifies how many positive predictions a model makes are correct, and recall quantifies how many total positive class instances in a dataset were correctly predicted as positive by a model. With an imbalanced dataset, the goal is to improve recall without significantly hurting precision. When working with imbalanced data, it is more appropriate to try and maximize precision when minimizing false positives and to maximize recall when minimizing false negatives (He and Ma 2013). The F1 measure balances precision and recall by taking the harmonic mean of the two metrics and maximizing the F1-score, which allows us to maximize both precision and recall and is commonly used with imbalanced datasets. In this study, since we are more concerned with predicting as many HMs as possible and using the RF model to evaluate HM drivers, we are more interested in minimizing false negatives.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2(\frac{precision * recall}{precision + recall})$$

The precision-recall curve (PRC) visually summarizes the trade-off between the TP rate and the positive predictive value for a model using different probability thresholds (Saito & Rehmsmeier, 2015). Neither precision nor recall utilizes the TN metric and is more focused on accurately predicting the minority class. Therefore, PRC curves are not influenced by how many majority class instances are correctly identified and will not be skewed by imbalanced classifiers' strong performance in predicting majority classes. In this case, the PRC curve will not be affected by the RF's bias toward predicting non-HMs correctly and will instead hone in on the RF's ability to correctly predict HMs (Saito & Rehmsmeier, 2015). A no-skill classifier (a classifier that cannot discriminate between classes) is represented on a PRC as a horizontal line whose position changes based on the class distribution of the data. This baseline is calculated as the ratio of positive cases to the whole dataset, and for a perfectly balanced class, would be 0.5. A perfect classifier would be two straight lines, and a robust model would be a curve whose convex hull is toward the (1,1) point. As a model's performance weakens, the shape bows towards the no-skill line, with the convex hull pointing towards (0,0). The area under the PRC is also used as a metric for the model. The AUC-PR is calculated as the trapezoidal area under the curve, and in a perfect model, it will have a value of 1 (Davis & Goodrich, 2004).

2.6 RF Feature Importance

For each RF in this study, we examined the importance of various potential environmental predictors to determine which variables were most impactful in HM classification. Features in machine learning refer to the variables fed into the model as input that it uses for predictions (Zvornicanin et al., 2023). The default technique for assessing the importance of each variable in an RF is impunity-based importance. In impunity-based approaches, the significance of a variable is determined by how much it can reduce impunity and improve prediction accuracy, quantified as the mean decrease in impunity (Par et al., 2018). Variables that can create clear, pure splits in a tree, leading to cleaner data segmentation, will be given higher importance in an impunity-based approach. However, the default impunity-based approach can often be misleading because the variable analysis is performed on training data and is vulnerable to effects from model overfitting. Impunity-based approaches are especially unsuitable for datasets with high cardinality (Stobhl et al., 2007; Parr et al., 2018). The consensus is that it is most appropriate to calculate the importance of variables using a permutation algorithm, which is a more reliable, model-agnostic approach (Breiman, 2001; Stobhl et al., 2007; Parr et al., 2018). The importance of the permutation feature is incredibly insightful when relationships are non-linear (Breiman, 2001). Permutation feature importance computes the importance of each

variable by 1) calculating a baseline accuracy score for RF using the test dataset, 2) randomly shuffling measurements in a single feature and then re-testing the RF with the perturbed sample set and calculating the new accuracy score, 3) the importance of the shuffled feature is determined by the drop in accuracy between the baseline and perturbed RFs (Parr et al., 2018). The higher the drop in accuracy between the baseline and shuffled models, the higher the importance of the variable. This workflow is repeated for every variable selected in the RF. While this process is computationally intensive, it yields a more accurate representation of the importance of the different variables tested.

Our initial logistic regression results for identifying HMs indicated non-linear relationships between the selected variables (Table 1) and HMs and CH4 emissions in general in this data. Because the wetland HM data exhibits high cardinality and non-linear relationships, using permutation feature importance instead of standard feature importances is a more suitable choice for assessing the impact of the variables on RF model predictions. We used <Sklearn> in Python to rank the features in RF by importance score.

3.0 Results

We built four RFs in this study: an unbalanced RF trained with the original

ratio of HM to non-HM, a balanced RF trained with the randomly undersampled training data, a balanced RF with data only from tidal sites, and a balanced RF with data only from non-tidal sites. The unbalanced RF was built using default hyperparameters with bootstrap sampling, 100 trees, a minimum sample split of two, a minimum leaf side of one, and the square root of the total number of features determined no maximum tree depth and the number of features. Using the K-fold cross-validation algorithm, we found that the highest accuracy for our balanced RF was achieved when the number of features to consider for 'best split' was calculated as the square root of total samples and tree growth did not use bootstrap sampling in the combined RF, tidal RF, and the non-tidal RF. The tree and forest hyperparameters that yielded the highest accuracy model for the balanced RF were 600 trees, the minimum number of samples required to split an internal node was three, the minimal number of samples required to be a terminal leaf node was one, and the maximum tree depth was 100. For the balanced non-tidal RF, we identified the best number of 500 trees, a minimum sample split of three, a minimum sample leaf of two, and a maximum tree depth of 70. The tidal RF achieved the highest accuracy with 1000 trees, a minimum sample split of two, a minimum leaf size of one, and a maximum tree depth of 40.

3.1 Unbalanced Random Forest Performance

To highlight the importance of addressing imbalanced training data when using RFs, we built an RF trained with the full unbalanced HM dataset. We used this model's performance as a baseline to compare the performance of the balanced RFs. The unbalanced RF's accuracy was exceedingly high at 99.7%, indicating very few prediction errors relative to the total number of predictions made. However, this accuracy metric is heavily skewed by correct non-HM predictions and obscures the model's inability to predict HMs correctly. The balanced accuracy presented in Table 3 provides a more authentic measure of the model's ability to predict HM and non-HMs. The unbalanced RF's confusion matrix also illustrates this model's ineptitude in predicting actual HMs - identifying only 3.24% of true HMs in the test dataset. Because HMs are exceedingly rare compared to non-HM in the unbalanced training dataset, this model did not have enough exposure to HM and could not correctly learn the decision boundary between the two classes.

Table 6: RF Performance Comparison

Metric	Unbalanced RF	Balanced RF	Balanced RF Non-Tidal	Balanced RF Tidal
Balanced Accuracy	51.6	77.18	78.2	71.32
Precision	0.763	0.079	0.075	0.071
Recall	0.032	0.77	0.765	0.694
F1	0.062	0.144	0.136	0.129

AUC-ROC	0.85	0.85	0.86	0.79
AUC-PRC	0.26	0.19	0.19	0.14

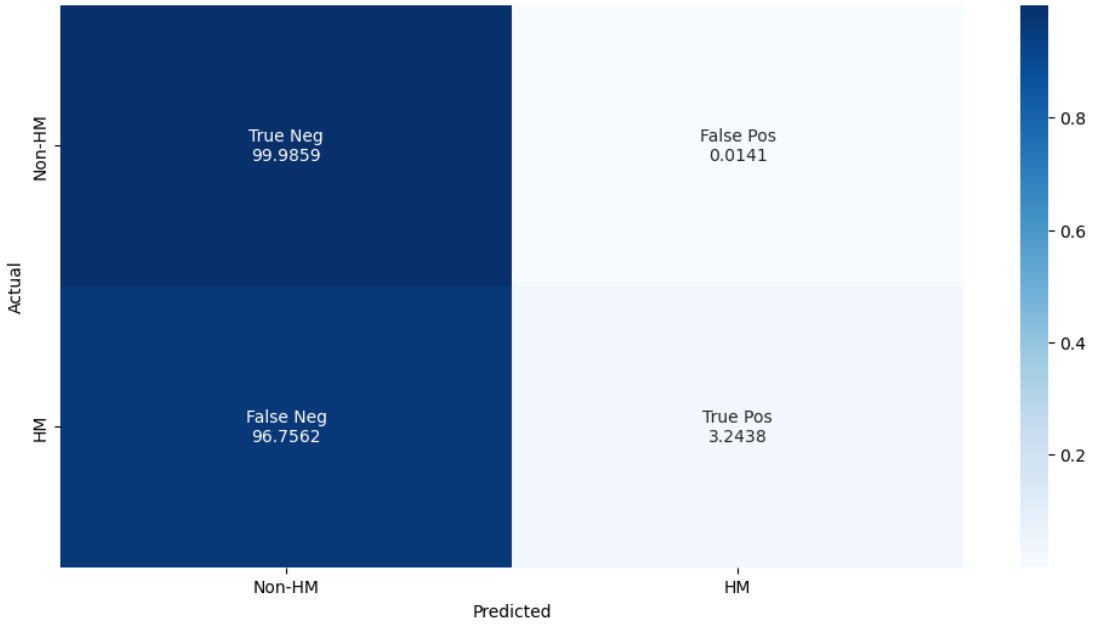
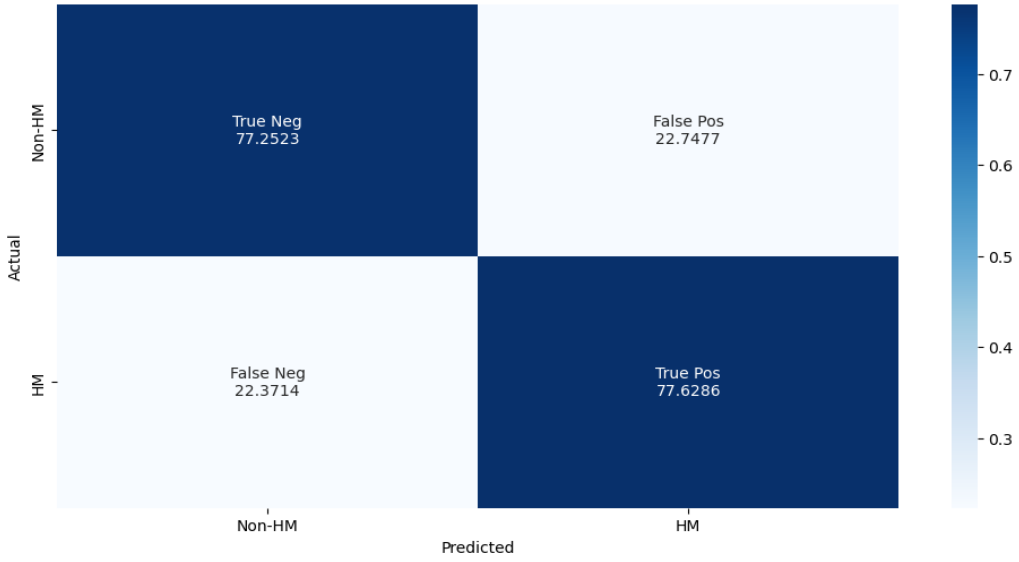


Figure 17: Unbalanced RF Confusion Matrix. Displays true, false, positive, and negative rates in the respective quadrants.

3.2 Optimized Combined Random Forest Results

The optimized RF trained with the randomly undersampled, balanced HM dataset outperformed the unbalanced RF and could identify true HMs faster. A comparison of the out-of-box balanced RF and the optimized balanced RF is

presented in Supplementary Materials. The performance metrics of the balanced RF are presented in Table 3. The balanced accuracy of the balanced RF improved to 77.18% compared to the 51.6% balanced accuracy in the unbalanced RF. While the overall accuracy of the balanced RF is lower than the unbalanced RF, we found this model was more appropriate for our use case because it can more accurately identify HM, which is our class of interest. The confusion matrix highlights the balanced RF's improved ability to detect true HMs at 77.6% and true non-HMs at 77.2%. It is worth noting that the FPR in the balanced RF increased compared to the FPR unbalanced RF, indicating that the model predicts too many HM.



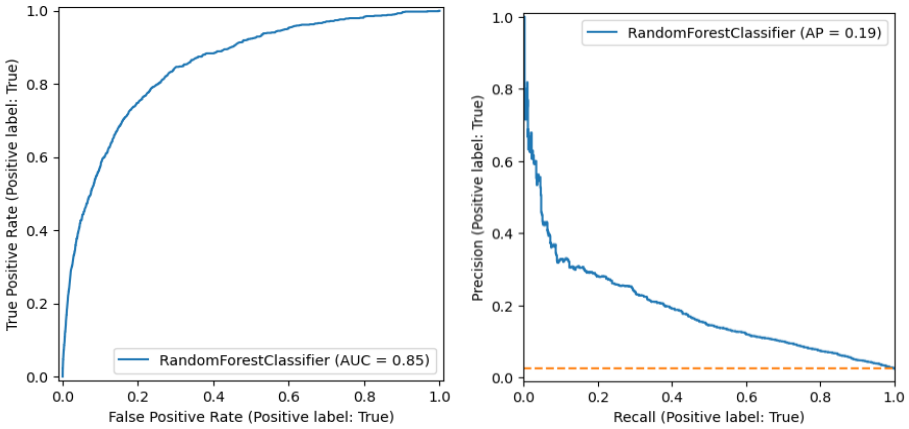


Figure 18: Balanced RF Confusion Matrix, ROC, and PR-Curve

The trade-off between TPR and FPR is further illustrated in the ROC Curve in Figure 18, which shows the rates across different classification thresholds. The no-skill classifier line is shown as the gray one-to-one dotted line, and the peak model performance region in the ROC plot is at (0,1), where TPR is 100%, and FPR is 0%. The shape of the ROC curve for the balanced RF indicates the model is performing well, and the AUC of 0.86 further establishes the model's strong performance. The ROC Curve reflects that the balanced RF can identify true HMs 78% of the time and only mislabel non-HMs as HMs 22% of the time (also seen in the confusion matrix). However, as noted by Saito and Rehmsmeier (2015), when ROC Curves and AUC are used with highly imbalanced data, they can present an overly optimistic assessment of model performance, and we determined the ROC and AUC for the balanced RF presented here exhibit this overly optimistic assessment of model

performance. The optimistic bias in the balanced RF's ROC results emerges because the FPR diminishes when the volume of actual negatives is present, as in our test dataset, where there are 33,490 non-HMs and 791 HMs. Consequently, the substantial number of false positives (8062) exerted minimal influence on the FPR, resulting in an elevated AUC value that failed to capture the false positive problem with the balanced RF model.

To get a more straightforward look at the model assessment, we turned to the PR Curve presented in Figure 18, which focuses on the predictions of the minority class. The no-skill classifier (where all predictions would be for the positive class) line is shown as a dotted horizontal line. The shape of the PR Curve and the AUC of 0.19 for the balanced RF denote a fair model performance. The balanced RF model has a high recall, which denotes that the model correctly predicted the majority of the HM instances in the test dataset. However, per the equation for precision presented above, a large number of false positives relative to the number of true positives will result in a low precision. While the balanced RF model correctly identified 77% of the true HMs in the test dataset, it also incorrectly flagged 22% of the non-HM as HM. These false positives dropped the precision score, and when the model was assessed across different classification thresholds to make the PR curve, it yielded a PR curve shape that indicates fair to poor performance. The maximization of recall vs precision scores seemingly flipped between the balanced and unbalanced RFs. In the

unbalanced RF, there was high precision and low recall, indicating that the model excelled at classification overall but did not perform well for minority class predictions. The balanced RF, on the other hand, had low precision and high recall scores, indicating that the model identifies most of the actual positive instances, but it is prone to false positives.

We computed the permutation feature importances for the balanced RF to assess which EC variables are most valuable when predicting HM of FCH4, which are presented in Figure 19. Negative values in the balanced RF's permutation feature importances indicate that permuting the values of a particular variable improved the model's accuracy, which implies that perturbing its values helps reduce noise or overfitting. The most valuable parameters for predicting HM were latent heat exchange, wind speed, wind direction, and photosynthetic photon flux density.

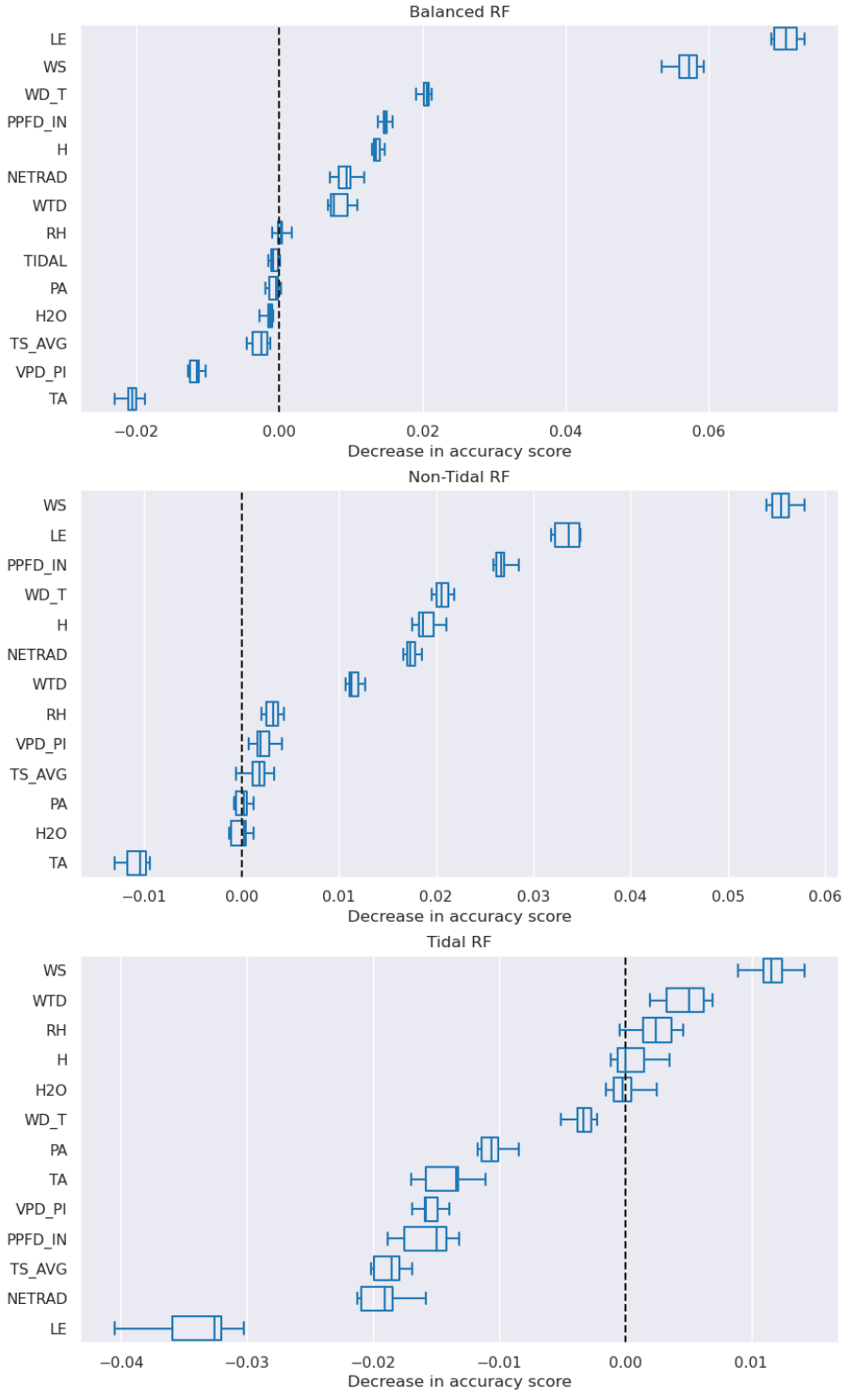
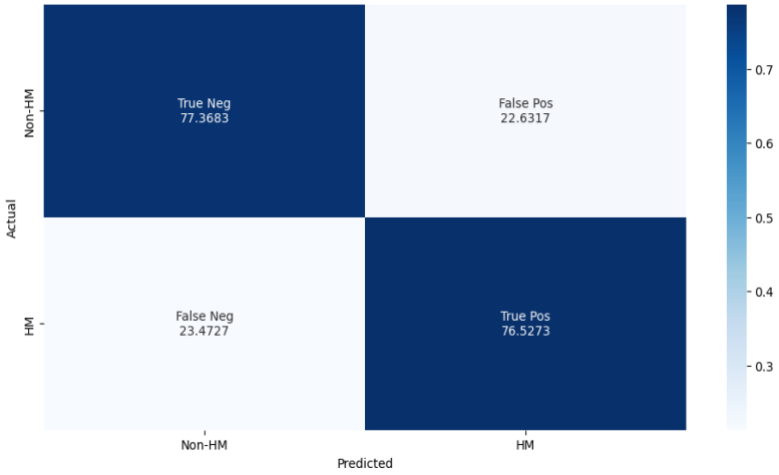

Balanced RF
LE
WS
WD_T
PPFD_IN
H
NETRAD
WTD
RH
TIDAL
PA
H2O
TS_AVG
VPD_PI
TA
−0.02
0.00
0.02
0.04
0.06
Decrease in accuracy score
Non-Tidal RF
WS
LE
PPFD_IN
WD_T
H
NETRAD
WTD
RH
VPD_PI
TS_AVG
PA
H2O
TA
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Decrease in accuracy score
Tidal RF
WS
WTD
RH
H
H2O
WD_T
PA
TA
VPD_PI
PPFD_IN
TS_AVG
NETRAD
LE
−0.04
−0.03
−0.02
−0.01
0.00
0.01
Decrease in accuracy score

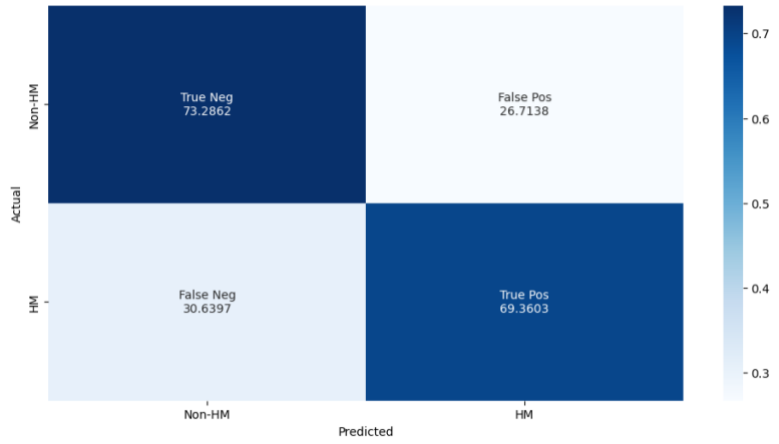
Figure 19: Feature Importances in Balanced RF, Non-Tidal RF, and Tidal RF.

Based on the feature importances, we were curious if the class imbalance drove the sub-par performance of the balanced RF or if combining tidal and non-tidal sites into one model was complicating the RF's ability to correctly identify HM because the hydrology and CH4 pathways and drivers are different between tidal and non-tidally influenced wetlands. To address this question, we created separate models for the tidal and non-tidal sites and generated a new set of predicted HM with the new tidal and non-tidal models.

3.2 Non-Tidal Random Forest Results

The non-tidal balanced RF was trained with US-TW1, US-TW4, US-TW5, US-MYB, and US-SNE data. We built a model with the combined non-tidal site data and optimized it using the K-fold cross-validation technique (comparison of base to optimized model can be found in Supplementary Materials). The standard and balanced accuracy of the non-tidal RF was similar to the combined RF's, with both accuracies hovering around 78% for the non-tidal RF (Table 4). Looking at the confusion matrix for the non-tidal RF in Figure 20 we can see that the TPR is high, and the model can identify true HM 76.5% of the time. We also see the same high FPR rate in the non-tidal model, with 22% of non-HMs incorrectly classified as HMs.


True Neg
77.3683
False Pos
22.6317
False Neg
23.4727
True Pos
76.5273
Non-HM
HM
Actual
Non-HM
HM
Predicted
0.7
0.6
0.5
0.4
0.3


True Neg
73.2862
False Pos
26.7138
False Neg
30.6397
True Pos
69.3603
Non-HM
HM
Actual
Non-HM
HM
Predicted
0.7
0.6
0.5
0.4
0.3

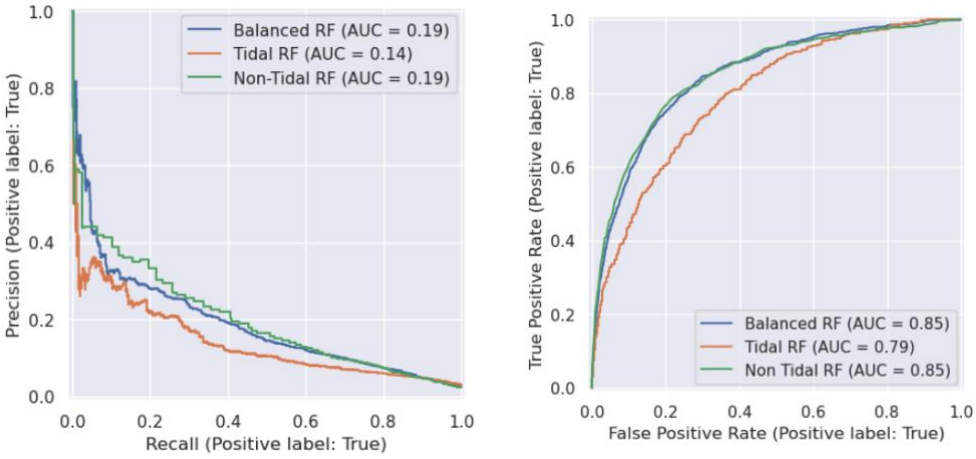


Figure 20: Tidal and Non-Tidal Confusion Matrices, ROC, and PR-Curves

The ROC and PR Curves display a similar model performance to the combined RF. Again, the ROC Curve likely gives an optimistic assessment of the non-tidal RF's performance driven by the high TPR, high TNR, and the small number of false positive predictions relative to the total number of true negatives. The PR Curve for the non-tidal RF has only a marginally lower AUC than the combined RF, reflecting that this model also has high recall and low precision metrics. However, as shown in Table 1, the non-tidal RF has slightly higher precision and slightly lower recall than the combined RF, resulting in a slightly higher F1-Score, indicating a marginally more balanced tradeoff between precision and recall.

Interestingly, the permutation feature importances for the non-tidal RF were similar to those of the combined RF. The three most important features in the

combined and non-tidal RF were wind speed, latent heat exchange, and sensible heat exchange. However, in the non-tidal RF, wind speed outranked latent heat flux. The fourth, fifth, and sixth most important features were photosynthetic photon density flux, wind direction, and net radiation, and in the combined RF, these features were also reasonably important to model accuracy. The feature importance of the non-tidal RF likely mirrors the combined RF so closely because there is more data from non-tidal sites than tidal sites in the combined RF, so they exert more influence on model training and characteristics.

3.3 Tidal Random Forest Results

There was a significant drop in model performance between the tidal, non-tidal, and combined models. A caveat about the tidal RF's performance is that it was trained with significantly fewer data (n= 33,480) than the combined (n=181,672) and non-tidal (n=148,192) RFs. The tidal RF's accuracy and balanced accuracy were 73.17% and 71.32%, respectively. The confusion matrix for the tidal RF in Figure 20 shows that the TPR is lower than that of the combined and non-tidal RFs at 69.4%. The FPR rate for the tidal RF is also significantly higher than the other RFs at 26.7%. The FPR and TPR indicate a significant drop in performance relative to the combined and non-tidal RFs.

The poor performance of the tidal RF is reflected in the ROC and PR curves in Figure 20. The ROC Curve shape has bowed, and the AUROC has dropped to 0.79 for the tidal curve. The ROC Curve's shape is likely influenced by the fewer true negatives flagged by the tidal RF. The PR curve has a significantly lower AUC than the combined and non-tidal RFs, and the PR curve's shape hovers closer to the no-skill line. This cure shape indicates that this model's ratio between precision and recall is even more skewed. The imbalance in precision and recall is seen in the F1 score in Table 1, which is lower for the tidal RF than the score in the combined and non-tidal RFs. The feature importances for the tidal RF differ more substantially from those of the combined RF feature importances than those of the non-tidal importances. The most important feature was still wind speed, but the rest of the feature importances were shuffled around. Notably, latent heat, which was so valuable in the combined and non-tidal RFs, is the second to least important feature in the tidal RF. Relative humidity and water vapor flux were much more influential on model performance in the tidal RF than in the combined and non-tidal RFs.

Discussion

4.1 Benefits of Balancing Data in Biogeochemical Modeling

In this study, we used several RF models to assess the predictability of HMs

of FCH4 in the SF Bay-Delta. The RF was envisioned as a model that could potentially predict the occurrence of HMs across the Bay-Delta at sites not monitored by EC towers. We evaluated the performance of an RF trained with unbalanced training data vs. an RF trained with balanced data. Balancing data is a critical first step when building an RF that can accurately predict HM, and we found that the balanced accuracy of the unbalanced RF improved from 51.6% to 77.18% when we balanced the RF training data. Exposing our RF to more HM during training removed the unbalanced RF's bias towards only correctly identifying the majority class non-HMs. With the balanced training data, we achieved a 78% TPR, a significant improvement from the unbalanced 3.24% TPR. Balancing datasets with severe class imbalance is commonplace in machine learning and medical fields, and class imbalance is becoming more well-researched in the context of biogeochemistry and ecology (Wilson et al., 2020; Salas-Eljatib et al., 2018; Bekendorf et al., 2023; Bourel et al., 2021). In biogeochemistry, numerous phenomena of interest are stochastic; thus, many datasets are imbalanced. As data-driven approaches become increasingly popular in the biogeochemical field, machine learning practitioners should ensure that class imbalance does not interfere with model performance. Addressing imbalance training data is especially important for researchers hoping to model HSHM, where the class imbalance is extreme. The results of our balanced RF compared to the unbalanced RF demonstrate the importance of balancing data before trying to model the minority class.

4.2 Balanced RF Interpretation and Limitations

The RF fitted with balanced data from tidal and non-tidal sites was used to assess the potential for forecasting HM of FCH4 in the Bay-Delta region. The optimized RF could predict the absence or presence of HMs with an overall 77% accuracy. The balanced RF outperformed the balanced logistic regression, CART, and C4.5 classifiers we tested and correctly identified more HMs than the baseline suite of models. Compared to our baseline suite of models, this model's performance indicates that RFs and other ensemble-supervised learning models hold potential for predicting extreme events in biogeochemical and ecological studies. However, while the balanced model could correctly identify 78% of the true HMs and 77% of the true non-HMs, it also had a high false positive rate, which we identified as the most significant shortcoming in model performance. Overall, the balanced RF tended to overpredict HM at every site. We attribute this over-prediction to the class imbalance problem. In the Supplementary Materials, we show that out of the different resampling techniques and ratios we tested, a randomly undersampled 1:1 balanced dataset was ideal for training the RF and maximizing true HM identification. However, we suspect that even with the optimal balancing, the dataset removed too many useful measurements that would have helped the RF learn the decision boundary between HM and non-HM better. Therefore, when the balanced RF was

presented with the unseen test data, the model could not always distinguish between true HM and non-HMs. We explored the possibility that underlying mechanistic differences between sites influence model performance while maintaining a big data approach by assessing the split tidal and non-tidal models. The lackluster results of these models allowed us to rule out the idea that the balanced RF was underperforming because we tried to predict tidal and non-tidal HM dynamics in one combined model. Instead, we determined that the underlying class imbalance was likely caused by the high FPR rate and not by an HM mechanistic-driven issue.

The balanced RF could be improved by including EC flux variables in the model, such as FC and the Monin–Obukhov length, which relates to the EC tower flux footprint. However, we excluded any EC-specific variables in this study to assess the potential for scaling the RF to sites without EC monitoring. We limited the predictor variables to those measured by standard environmental monitoring systems or remote sensing. We posit that including ecosystem productivity measures, such as GPP, RECO, and NEE, might improve the predictive capability of the balanced RF since there is a close tie between plant productivity and methane production and emission. Here, we did not include these productivity metrics because they were unavailable at all nine sites. However, they hold potential for upscaling because GPP and NEE can be estimated from remote sensing and be used at wetlands that do not have EC towers. Results of the RF trained with all available EC data are presented in

Supplementary Materials. We also hypothesize that using Feature Forward Selection to build the RF might improve performance and training computation time. Forward Feature Selection enhances model accuracy by iteratively selecting informative features, thus improving model interpretability and reducing dimensionality. Feature Forward Selection has been used in EC data CH4 upscaling models and has proven useful (McNicol et al., 2023).

We extensively tested different resampling and class weighting techniques to deal with imbalanced data classification, which can be found in Supplementary Material. Future work to improve the balanced RF presented here could include incorporating uncertainties or model-level classification threshold adjustments to create a more discerning model. RFs are non-parametric ensemble models, and prediction errors are not quantified directly because classification is based on tree majority voting. Computing uncertainty for the balanced RF predictions would help us understand the model's confidence in each prediction and could shed light on why many non-HMs are flagged as HMs. Recent efforts in machine learning have sought to add an uncertainty component to RFs using probabilistic nodes, Monte Carlo techniques, confidence intervals, and hypothesis testing (Bauman et al., 2015; Coulston et al., 2016; Mentch & Hooker, 2016). RFs are not pattern recognition algorithms and cannot link model decision-making to physical and dynamic components in a system like certain deep learning algorithms can achieve (Giamalaki,

2021). Therefore, a more sophisticated machine learning algorithm capable of true pattern recognition, like a neural network or fuzzy model approach, could improve this study's HM predictions of the balanced RF.

HSHMs are generally understudied and poorly understood across ecosystems. They are usually used as a framework to characterize dynamics in a system and not as an extreme value to model and predict. The results of the balanced RF represent a new effort to use data-driven models to model HM. While the relatively high FPR in the balanced RF is a drawback of the balanced RF, the overall model accuracy shows promise in predicting the occurrence of HM. Additionally, the feature importances in the balanced, tidal, and non-tidal RFs provide helpful information about FCH4 HM predictors and indicate a strong potential for upscaling across sites.

4.3 Feature Importance and Model Applications

While the balanced RF is not a perfect classifier, we can still use the importance of environmental variables to glean information about FCH4 HM predictors that wetland managers and future modelers could use. Note that variable importance does not reflect the intrinsic predictive value of a variable but rather how important this variable is for this specific model, and we cannot conclusively call the identified important variables in the RF FCH4 HM mechanistic drivers. However, we can compare the most essential variable identified by the model with known

biogeochemical and micrometeorological variables that influence and correlate with FCH4. The most important variables in the balanced RF, latent and sensible heat exchange, wind speed and direction, photosynthetically active photon flux (PPFD), net radiation, and water table depth align with our biogeochemical understanding of CH4 emission pathways. This alignment lends credibility to the importance of variable rankings and the model results, as we would expect the variables observed to drive CH4 emissions to be most beneficial to our RF when predicting HM of FCH4. Methanogenesis in wetlands occurs in the anaerobic subsurface as microbes in the soil break down organic matter and produce CH4. This microbially generated CH4 is stored in the soil, water, or vegetation roots. From these reservoirs, wetland CH4 can follow through three pathways to the atmosphere: gas bubble ebullition, diffusion from wetland soil and water, and plant-mediated transport (Turner et al., 2020). In ebullition, CH4 moves directly from the water columns into the atmosphere via gas bubbles rising to the surface. CH4 can escape from wetland soil and open water through diffusion at the air/water and air/soil interfaces. Plant-mediated CH4 transport occurs when plants remove CH4 from the soil and move it to the atmosphere via the spongy aerenchyma tissue in their roots and stomata (Villa et al., 2020). Wind speed and latent heat exchange impact the CH4 concentration gradient between soil or water and the atmosphere, impacting diffusive gas fluxes across these air-water-soil boundary layers.

Latent heat exchange in wetlands has been linked to methane emissions as an influencing factor in several studies (Li et al., 2023; Villa et al., 2020; Morin et al., 2014). It is, therefore, unsurprising that latent heat was the most important variable in the balanced RF. Morin et al. 2014, make a crucial distinction that latent heat fluxes are more likely to be a coincident rather than a directly driving variable of FCH4. Latent heat fluxes can be conceptualized as water fluxes and are related to CH4 emissions through plants and evaporation in a non-linear relationship (Morin et al., 2014). Latent heat can increase plant productivity and accelerate the emission through plant-mediated methane pathways and the methane emission rate from unsaturated soil. The RFs are likely learning this non-linear relationship between latent heat flux and methane emissions and using it to help predict the occurrence of HM of FCH4.

Wind speed was the second most important variable in the balanced RF and the most important variable in the tidal and non-tidal RFs and related to the diffusion and ebullition CH4 pathways. Wind speed is a physical micrometeorological measurement with strong influences on gas fluxes, including FCH4, and it has been shown that both diffusion and ebullition can increase with wind speed. Mechanistically, wind speed influences CH4 flux by moving air parcels at the air-water and air-soil interface and changing the micrometeorological pressure conditions such that CH4 in the wetland soil or water can overcome the pressure differential and escape to the atmosphere and transmitting turbulent energy that perturbs sediment

structure and initiates bubble release (Keller & Stallard, 1994) as well as removing stagnant layers and increasing diffusive gradients. Ebullition is a distinctly episodic process and is likely a significant pathway for HMs of FCH4 at coast wetlands. Therefore, we would also expect it to be a valuable predictor for HM of FCH4 in our RFs.

Incoming Photosynthetic Photon Flux Density (PPFD) was also a high-ranking feature in the balanced, tidal, and non-tidal RFs. PPFD photon flux density of photosynthetically active radiation (PAR) measures the number of photons in the 400- to 700-nm waveband per unit of time on a unit surface (Rabinowitz & Vogel, 2009). PAR is required for photosynthesis and plant productivity, and higher PAR promotes plant growth. The RFs are likely learning the relationship between increased PAR and increased plant productivity and the relationship between plant productivity and CH4 emissions. Vegetated soils can exhibit elevated CH4 fluxes compared to non-vegetated soils because plants contribute accessible carbon substrate used by methanogens (Shahan et al., 2021; Bansal et al., 2020; Hatala et al., 2012; Reid et al., 2013; Rey-Sanchez et al., 2018). Seasonal variations in FCH4 at the highly vegetated sites in this study (MYB, TW1, TW4, TW5, SNE, DMG) move in concert with the growing season. Indeed, in the Delta, CH4 emissions are highest during the growing season when it is warm, the CH4-producing microbes are the most active, and plant productivity peaks as they deposit lots of organic matter into

the soil (Bhullar et al., 2013).

Water table depth is firmly grounded in the literature as a control on FCH4, particularly at tidally influenced sites (Bhullar et al., 2013; Evans et al., 2021). The relationship between the water table depth and FCH4 is complicated and difficult to isolate, as many different factors in a wetland system can influence CH4 emissions. Several studies have found that CH4 ebullition can occur at tidal sites when hydrostatic pressure in the sediment is released during low tide (Keller & Stallard, 1994; Schmid et al., 2017). Water table depth has also been shown to strongly shape the size and diversity of the CH4-producing microbe population at peatland (Tian et al., 2023). Moore and Knowles (1989) found that CH4 evolution decreased logarithmically as the water table depth lowered, which is not unexpected as water-table depth is linked to increased oxidizing conditions, resulting in reduced emissions from unsaturated soils (Grunfeld & Brix, 1999). Generally, when water table depth is high, wetland soils are saturated, anaerobic conditions ideal for methanogenesis persist, and more CH4 is available for emission (Evans et al., 2021; Peacock et al., 2024). Fluctuations in water-table depth, therefore, can inhibit or promote FCH4 and influence the occurrence of HM. For these reasons, we expected the water table depth parameter to be more beneficial for HM identification than it actually was. We hypothesized it would be particularly useful for the tidal RF. However, the water table depth was a mid-ranking feature in the balanced RF and tidal RF, which

suggests that while water table depth might influence overall FCH4 dynamics, it does not predict anomalously high CH4 emissions well.

The valuable predictors identified by the balanced RF present an opportunity to examine further what drives HM of FCH4 and pose a question of whether or not the known influences on baseline FCH4 still apply to instances of extreme flux or if they are driven by different mechanisms entirely. Additionally, the predictors identified in this model present an opportunity for upscaling as they can be measured without EC towers. Therefore, if we were to compile data from weather stations and remote sensing for each of the predictors in the RF from a site without an EC tower, we could use a model with this new data to attempt to predict the occurrence of HM at a new site.

RF classifiers provide a simple, data-driven approach to unraveling the complexities of what triggers HM of FCH4. The balanced RF presents a promising approach to HM prediction in wetland sites. HM predictability and a reliable near-real-time predictive model would prove valuable to the SF Bay-Delta restoration projects and wetland managers. Identifying what causes these highly impactful but temporally rare emission events would allow managers to make decisions regarding wetland hydrology, such as flooding or replanting historic vegetation species to a wetland, with knowledge of how these actions would influence short-term CH4

balance at a site. Additionally, the upscaling potential of our model would be helpful for wetland managers who do not have EC data for their site, as they could use standard environmental parameters as model input and generate predictions for HM occurrence. Restored and managed wetlands are often considered significant C storage ecosystems within the year-to-decade timelines. Honing in on the HSHM of FCH4 and developing techniques to minimize these disproportionate emission instances provides a unique intervention method for managers on the timescale relevant to day-to-day human activities. Focusing on HSHM allows managers to prioritize their time and resources for high-impact CH4 emission prevention measures.

The work presented here provides a strong case for the overall predictability of FCH4 HM presence/absence and upscaling using data-driven models. This study also illustrates the need for dataset balancing when working with inherently rare HM. Such FCH4 HM predictive models can also provide insight into what physical and biogeochemical characteristics of a wetland influence the occurrence of HM and provide valuable information for managers and future restoration work in the SF Bay-Delta and beyond. Future work to improve upon the balanced RF presented in this study should focus on reducing the number of false positives flagged by the model, adjusting model thresholding for identifying minority classes in an imbalanced dataset, and extrapolating the HM prediction capabilities of RFs to more sophisticated

machine learning models that can handle the complicated pattern recognition needed to parse the many influences on CH4 cycling in wetlands. Future work could also include developing predictive HM models for different regions and types of wetlands using the best predictors identified here as upscaling input.

References

Al-Haj AN, Fulweiler RW. A synthesis of methane emissions from shallow vegetated coastal ecosystems. Glob Change Biol. 2020; 26: 2988–3005. https://doi.org/10.1111/gcb.15046

Alongi, D. M. (2020). Carbon Balance in Salt Marsh and Mangrove Ecosystems: A Global Synthesis. *Journal of Marine Science and Engineering*, *8*(10), Article 10. https://doi.org/10.3390/jmse8100767

Anderegg, W. R. L., Schwalm, C., Biondi, F., Camarero, J. J., Koch, G., Litvak, M., Ogle, K., Shaw, J. D., Shevliakova, E., Williams, A. P., Wolf, A., Ziaco, E., & Pacala, S. (2015). Pervasive drought legacies in forest ecosystems and their implications for carbon cycle models. *Science*, *349*(6247), 528–532. https://doi.org/10.1126/science.aab1833

Anderson, F. E., Bergamaschi, B., Sturtevant, C., Knox, S., Hastings, L., Windham-Myers, L., Detto, M., Hestir, E. L., Drexler, J., Miller, R. L., Matthes, J. H., Verfaillie, J., Baldocchi, D., Snyder, R. L., & Fujii, R. (2016). Variation of energy and carbon fluxes from a restored temperate freshwater wetland and implications for carbon market verification protocols. *Journal of Geophysical Research: Biogeosciences*, *121*(3), 777–795. https://doi.org/10.1002/2015JG003083

Anthony, T. L., & Silver, W. L. (2021). Hot moments drive extreme nitrous oxide and methane emissions from agricultural peatlands. *Global Change Biology*, *27*(20), 5141–5153. https://doi.org/10.1111/gcb.15802

Arias-Ortiz, A., Oikawa, P. Y., Carlin, J., Masqué, P., Shahan, J., Kanneg, S., Paytan, A., & Baldocchi, D. D. (2021). Tidal and Nontidal Marsh Restoration: A Trade-Off Between Carbon Sequestration, Methane Emissions, and Soil Accretion. *Journal of Geophysical Research: Biogeosciences*, *126*(12), e2021JG006573. https://doi.org/10.1029/2021JG006573

Arora, B., Briggs, M. A., Zarnetske, J. P., Stegen, J., Gomez-Velez, J. D., Dwivedi, D., & Steefel, C. (2022). Hot Spots and Hot Moments in the Critical Zone: Identification of and Incorporation into Reactive Transport Models. In A. S. Wymore, W. H. Yang, W. L. Silver, W. H. McDowell, & J. Chorover (Eds.), *Biogeochemistry*

of the Critical Zone (pp. 9–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-95921-0_2

Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future: CARBON BALANCE and EDDY COVARIANCE. *Global Change Biology*, *9*(4), 479–492. https://doi.org/10.1046/j.1365-2486.2003.00629.x

Bange, H. W., Bartell, U. H., Rapsomanikis, S., & Andreae, M. O. (1994). Methane in the Baltic and North Seas and a reassessment of the marine emissions of methane. *Global Biogeochemical Cycles*, *8*(4), 465–480. https://doi.org/10.1029/94GB02181

Barnes, M. E., Johnson, D. W., & Hart, S. C. (2024). The Median Isn't the Message: Soil nutrient hot spots have a disproportionate influence on biogeochemical structure across years, seasons, and depths. *Biogeochemistry*, *167*(1), 75–95. https://doi.org/10.1007/s10533-023-01107-x

Batt, R. D., Carpenter, S. R., & Ives, A. R. (2017). Extreme events in lake ecosystem time series. *Limnology and Oceanography Letters*, *2*(3), 63–69. https://doi.org/10.1002/lol2.10037

Benhadi-Marín, J. (2018). A conceptual framework to deal with outliers in ecology. *Biodiversity and Conservation*, *27*(12), 3295–3300. https://doi.org/10.1007/s10531-018-1602-2

Benkendorf, D. J., Schwartz, S. D., Cutler, D. R., & Hawkins, C. P. (2023). Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models. *Ecological Modelling*, *483*, 110414. https://doi.org/10.1016/j.ecolmodel.2023.110414

Bernhardt, E. S., Blaszczak, J. R., Ficken, C. D., Fork, M. L., Kaiser, K. E., & Seybold, E. C. (2017). Control Points in Ecosystems: Moving Beyond the Hot Spot Hot Moment Concept. *Ecosystems*, *20*(4), 665–682. https://doi.org/10.1007/s10021-016-0103-y

Beyer, H. (1981). Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don

Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical Journal*, *23*(4), 413–414. https://doi.org/10.1002/bimj.4710230408

Bhullar, G. S., Edwards, P. J., & Olde Venterink, H. (2013). Variation in the plant-mediated methane transport and its importance for methane emission from intact wetland peat mesocosms. *Journal of Plant Ecology*, *6*(4), 298–304. https://doi.org/10.1093/jpe/rts045

Birkel, C., Soulsby, C., & Tetzlaff, D. (2014). Integrating parsimonious models of hydrological connectivity and soil biogeochemistry to simulate stream DOC dynamics. *Journal of Geophysical Research: Biogeosciences*, *119*(5), 1030–1047. https://doi.org/10.1002/2013JG002551

Bokde, N., Feijóo, A., & Kulat, K. (2018). Analysis of differencing and decomposition preprocessing methods for wind speed prediction. *Applied Soft Computing*, *71*, 926–938. https://doi.org/10.1016/j.asoc.2018.07.041

Bourel, M., Segura, A. M., Crisci, C., López, G., Sampognaro, L., Vidal, V., Kruk, C., Piccini, C., & Perera, G. (2021). Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Research*, *202*, 117450. https://doi.org/10.1016/j.watres.2021.117450

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2020). *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery. https://books.google.com/books?id=jaXJDwAAQBAJ

California Natural Resource Agency. (n.d.). *California EcoRestore Groundbreakings*. Retrieved March 18, 2024, from https://resources.ca.gov/Initiatives/California-EcoRestore/California-EcoRestore-Groundbreakings

Carpenter, S. R., Booth, E. G., Kucharik, C. J., & Lathrop, R. C. (2015). Extreme daily loads: Role in annual phosphorus input to a north temperate lake. *Aquatic Sciences*, *77*(1), 71–79. https://doi.org/10.1007/s00027-014-0364-5

Chen, C. (n.d.). *Using Random Forest to Learn Imbalanced Data*.

Chu, H., Christianson, D. S., Cheah, Y.-W., Pastorello, G., O'Brien, F., Geden, J., Ngo, S.-T., Hollowgrass, R., Leibowitz, K., Beekwilder, N. F., Sandesh, M., Dengel, S., Chan, S. W., Santos, A., Delwiche, K., Yi, K., Buechner, C., Baldocchi, D., Papale, D., … Torn, M. S. (2023). AmeriFlux BASE data pipeline to support network growth and data sharing. *Scientific Data*, *10*(1), 614. https://doi.org/10.1038/s41597-023-02531-2

Cook, T. (2016, March 15). *After a Century, Restored Wetlands May Still Be a Carbon Source*. Eos. http://eos.org/research-spotlights/after-a-century-restored-wetlands-may-still-be-a-carbon-source

Coulston, J. W., Blinn, C. E., Thomas, V. A., & Wynne, R. H. (2016). Approximating prediction uncertainty for random forest regression models. *Photogrammetric Engineering & Remote Sensing*, *82*, 189–197. https://doi.org/10.14358/PERS.82.3.189

Darrouzet-Nardi, A., & Bowman, William. D. (2011). Hot Spots of Inorganic Nitrogen Availability in an Alpine-Subalpine Ecosystem, Colorado Front Range. *Ecosystems*, *14*(5), 848–863.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. https://doi.org/10.1145/1143844.1143874

De Battisti, F., Porro, F., & Vernizzi, A. (2019). *The Gini Coefficient and the Case of Negative Values* (SSRN Scholarly Paper 3610829). https://papers.ssrn.com/abstract=3610829

Dye, D. G. (2004). Spectral composition and quanta-to-energy ratio of diffuse photosynthetically active radiation under diverse cloud conditions. *Journal of Geophysical Research: Atmospheres*, *109*(D10). https://doi.org/10.1029/2003JD004251

Ecoatlas. (n.d.). *Ecoatlas: Bay/Delta—Projects*. Retrieved March 18, 2024, from https://www.ecoatlas.org/regions/ecoregion/bay-delta/projects/10272

Evans, C. D., Peacock, M., Baird, A. J., Artz, R. R. E., Burden, A., Callaghan, N., Chapman, P. J., Cooper, H. M., Coyle, M., Craig, E., Cumming, A., Dixon, S., Gauci, V., Grayson, R. P., Helfter, C., Heppell, C. M., Holden, J., Jones, D. L., Kaduk, J., … Morrison, R. (2021). Overriding water table control on managed peatland greenhouse gas emissions. *Nature*, *593*(7860), 548–552. https://doi.org/10.1038/s41586-021-03523-1

Everitt, B. S., & Skrondal, A. (n.d.). *The Cambridge Dictionary of Statistics*.

Frolking, S., Roulet, N., & Fuglestvedt, J. (2006). How northern peatlands influence the Earth's radiative budget: Sustained methane emission versus sustained carbon sequestration. *Journal of Geophysical Research: Biogeosciences*, *111*(G1). https://doi.org/10.1029/2005JG000091

Frolking, S., & Roulet, N. T. (2007). Holocene radiative forcing impact of northern peatland carbon accumulation and methane emissions. *Global Change Biology*, *13*(5), 1079–1088. https://doi.org/10.1111/j.1365-2486.2007.01339.x

Frolking, S., Talbot, J., Jones, M. C., Treat, C. C., Kauffman, J. B., Tuittila, E.-S., & Roulet, N. (2011). Peatlands in the Earth's 21st century climate system. *Environmental Reviews*, *19*(NA), 371–396. https://doi.org/10.1139/a11-014

Gedney, N., Cox, P. M., & Huntingford, C. (2004). Climate feedback from wetland methane emissions. *Geophysical Research Letters*, *31*(20). https://doi.org/10.1029/2004GL020919

Giamalaki, K., Beaulieu, C., & Prochaska, J. X. (2022). Assessing Predictability of Marine Heatwaves With Random Forests. *Geophysical Research Letters*, *49*(23), e2022GL099069. https://doi.org/10.1029/2022GL099069

Golkar, F., & Shirvani, A. (2020). Spatial and temporal distribution and seasonal prediction of satellite measurement of CO2 concentration over Iran. *International Journal of Remote Sensing*, *41*(23), 8891–8909. https://doi.org/10.1080/01431161.2020.1788743

Griscom, B. W., Adams, J., Ellis, P. W., Houghton, R. A., Lomax, G., Miteva, D. A., Schlesinger, W. H., Shoch, D., Siikamäki, J. V., Smith, P., Woodbury, P., Zganjar,

C., Blackman, A., Campari, J., Conant, R. T., Delgado, C., Elias, P., Gopalakrishna, T., Hamsik, M. R., … Fargione, J. (2017). Natural climate solutions. *Proceedings of the National Academy of Sciences*, *114*(44), 11645–11650. https://doi.org/10.1073/pnas.1710465114

Grünfeld, S., & Brix, H. (1999). Methanogenesis and methane emissions: Effects of water table, substrate type and presence of *Phragmites australis*. *Aquatic Botany*, *64*(1), 63–75. https://doi.org/10.1016/S0304-3770(99)00010-8

Hagedorn, F., & Bellamy, P. (2011). Hot Spots and Hot Moments for Greenhouse Gas Emissions from Soils. *Soil Carbon in Sensitive European Ecosystems: From Science to Land Management*, 13–32. https://doi.org/10.1002/9781119970255.ch2

Hahn-Schöfl, M., Zak, D., Minke, M., Gelbrecht, J., Augustin, J., & Freibauer, A. (2011). Organic sediment formed during inundation of a degraded fen grassland emits large fluxes of CH_4 and CO_2. *Biogeosciences*, *8*(6), 1539–1550. https://doi.org/10.5194/bg-8-1539-2011

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, *42*(6), 1887–1896. https://doi.org/10.1214/aoms/1177693054

Harms, T. K., & Grimm, N. B. (2008). Hot spots and hot moments of carbon and nitrogen dynamics in a semiarid riparian zone. *Journal of Geophysical Research: Biogeosciences*, *113*(G1). https://doi.org/10.1029/2007JG000588

Hatala, J. A., Detto, M., Sonnentag, O., Deverel, S. J., Verfaillie, J., & Baldocchi, D. D. (2012). Greenhouse gas (CO2, CH4, H2O) fluxes from drained and flooded agricultural peatlands in the Sacramento-San Joaquin Delta. *Agriculture, Ecosystems & Environment*, *150*, 1–18. https://doi.org/10.1016/j.agee.2012.01.009

Hemes, K. S., Chamberlain, S. D., Eichelmann, E., Anthony, T., Valach, A., Kasak, K., Szutu, D., Verfaillie, J., Silver, W. L., & Baldocchi, D. D. (2019). Assessing the carbon and climate benefit of restoring degraded agricultural peat soils to managed wetlands. *Agricultural and Forest Meteorology*, *268*, 202–214. https://doi.org/10.1016/j.agrformet.2019.01.017

Hemes, K. S., Eichelmann, E., Chamberlain, S. D., Knox, S. H., Oikawa, P. Y.,

Sturtevant, C., Verfaillie, J., Szutu, D., & Baldocchi, D. D. (2018). A Unique Combination of Aerodynamic and Surface Properties Contribute to Surface Cooling in Restored Wetlands of the Sacramento-San Joaquin Delta, California. *Journal of Geophysical Research: Biogeosciences*, *123*(7), 2072–2090. https://doi.org/10.1029/2018JG004494

Herman, G. R., & Schumacher, R. S. (2018). Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Monthly Weather Review*, *146*(5), 1571–1600. https://doi.org/10.1175/MWR-D-17-0250.1

Hippke, M., David, T. J., Mulders, G. D., & Heller, R. (2019). Wōtan: Comprehensive Time-series Detrending in Python. *The Astronomical Journal*, *158*(4), 143. https://doi.org/10.3847/1538-3881/ab3984

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C. J., Benthuysen, J. A., Burrows, M. T., Donat, M. G., Feng, M., Holbrook, N. J., Moore, P. J., Scannell, H. A., Sen Gupta, A., & Wernberg, T. (2016). A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, *141*, 227–238. https://doi.org/10.1016/j.pocean.2015.12.014

Holm, G. O., Perez, B. C., McWhorter, D. E., Krauss, K. W., Johnson, D. J., Raynie, R. C., & Killebrew, C. J. (2016). Ecosystem Level Methane Fluxes from Tidal Freshwater and Brackish Marshes of the Mississippi River Delta: Implications for Coastal Wetland Carbon Projects. *Wetlands*, *36*(3), 401–413. https://doi.org/10.1007/s13157-016-0746-7

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal: EMJ*, *34*(6), 357–359. https://doi.org/10.1136/emermed-2017-206735

Huang, M., Wang, X., Keenan, T. F., & Piao, S. (2018). Drought timing influences the legacy of tree growth recovery. *Global Change Biology*, *24*(8), 3546–3559. https://doi.org/10.1111/gcb.14294

Huang, N., Wang, L., Zhang, Y., Gao, S., & Niu, Z. (2021). Estimating the Net Ecosystem Exchange at Global FLUXNET Sites Using a Random Forest Model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 9826–9836. https://doi.org/10.1109/JSTARS.2021.3114190

Hunt, R. J., Krabbenhoft, D. P., & Anderson, M. P. (1997). Assessing hydrogeochemical heterogeneity in natural and constructed wetlands. *Biogeochemistry*, *39*(3), 271–293. https://doi.org/10.1023/A:1005889319205

Ialongo, C. (2019). Confidence interval for quantiles and percentiles. *Biochemia Medica*, *29*(1), 010101. https://doi.org/10.11613/BM.2019.010101

IBM, C. (2024, February 29). *IBM Documentation: Modified Z-Score* [API]. https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-modified-z-score

Iglewicz, B., & Hoaglin, D. (2010). *How to Detect and Handle Outliers (e-book) | ASQ*. ASQC Quality Press. https://asq.org/quality-press/display-item?item=E0801

Jeffrey, L. C., Maher, D. T., Johnston, S. G., Kelaher, B. P., Steven, A., & Tait, D. R. (2019). Wetland methane emissions dominated by plant-mediated fluxes: Contrasting emissions pathways and seasons within a shallow freshwater subtropical wetland. *Limnology and Oceanography*, *64*(5), 1895–1912. https://doi.org/10.1002/lno.11158

Johnson, D. W., Glass, D. W., Murphy, J. D., Stein, C. M., & Miller, W. W. (2010). Nutrient hot spots in some sierra Nevada forest soils. *Biogeochemistry*, *101*(1), 93–103. https://doi.org/10.1007/s10533-010-9423-8

Kannenberg, S. A., Bowling, D. R., & Anderegg, W. R. L. (2020). Hot moments in ecosystem fluxes: High GPP anomalies exert outsized influence on the carbon cycle and are differentially driven by moisture availability across biomes. *Environmental Research Letters*, *15*(5), 054004. https://doi.org/10.1088/1748-9326/ab7b97

Kannenberg, S. A., Maxwell, J. T., Pederson, N., D'Orangeville, L., Ficklin, D. L., & Phillips, R. P. (2019). Drought legacies are dependent on water table depth, wood anatomy and drought timing across the eastern US. *Ecology Letters*, *22*(1), 119–127. https://doi.org/10.1111/ele.13173

Keller, M., & Stallard, F. (1994). Methane emission by bubbling from Gatun Lake, Panama. *Journal of Geophysical Research-Atmospheres*, 8307–8319. https://doi.org/10.1029/92JD02170

Knox, S. H., Jackson, R. B., Poulter, B., McNicol, G., Fluet-Chouinard, E., Zhang, Z.,

Hugelius, G., Bousquet, P., Canadell, J. G., Saunois, M., Papale, D., Chu, H., Keenan, T. F., Baldocchi, D., Torn, M. S., Mammarella, I., Trotta, C., Aurela, M., Bohrer, G., … Zona, D. (2019). FLUXNET-CH4 Synthesis Activity: Objectives, Observations, and Future Directions. *Bulletin of the American Meteorological Society*, *100*(12), 2607–2632. https://doi.org/10.1175/BAMS-D-18-0268.1

Knox, S. H., Sturtevant, C., Matthes, J. H., Koteen, L., Verfaillie, J., & Baldocchi, D. (2015). Agricultural peatland restoration: Effects of land-use change on greenhouse gas (CO2 and CH4) fluxes in the Sacramento-San Joaquin Delta. *Global Change Biology*, *21*(2), 750–765. https://doi.org/10.1111/gcb.12745

Kolus, H. R., Huntzinger, D. N., Schwalm, C. R., Fisher, J. B., McKay, N., Fang, Y., Michalak, A. M., Schaefer, K., Wei, Y., Poulter, B., Mao, J., Parazoo, N. C., & Shi, X. (2019). Land carbon models underestimate the severity and duration of drought's impact on plant productivity. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-39373-1

Koweek, D. A., Dunbar, R. B., Monismith, S. G., Mucciarone, D. A., Woodson, C. B., & Samuel, L. (2015). High-resolution physical and biogeochemical variability from a shallow back reef on Ofu, American Samoa: An end-member perspective. *Coral Reefs*, *34*(3), 979–991. https://doi.org/10.1007/s00338-015-1308-9

Krause, J. (n.d.). *Self-monitoring report—Eden Landing—2020 | South Bay Salt Ponds*. Self-Monitoring Report - Eden Landing - 2020. Retrieved March 18, 2024, from https://www.southbayrestoration.org/document/self-monitoring-report-eden-landing-2020

Langford, E. (2006). Quartiles in Elementary Statistics. *Journal of Statistics Education*, *14*(3). https://doi.org/10.1080/10691898.2006.11910589

Li, J., Hao, T., Yang, M., Chen, Z., Zhu, J., Wang, Q., & Yu, G. (2023). *Analysis of Methane Emission Characteristics and Environmental Response in Natural Wetlands* (SSRN Scholarly Paper 4611970). https://doi.org/10.2139/ssrn.4611970

Li, Y., Liu, M., & Wu, X. (2022). Reclaimed Water Reuse for Groundwater Recharge: A Review of Hot Spots and Hot Moments in the Hyporheic Zone. *Water*, *14*(12), Article 12. https://doi.org/10.3390/w14121936

Lin, W.-C., Tsai, C.-F., Hu, Y.-H., & Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, *409–410*, 17–26. https://doi.org/10.1016/j.ins.2017.05.008

Liu, L., Wang, D., Chen, S., Yu, Z., Xu, Y., Li, Y., Ge, Z., & Chen, Z. (2019). Methane Emissions from Estuarine Coastal Wetlands: Implications for Global Change Effect. *Soil Science Society of America Journal*, *83*(5), 1368–1377. https://doi.org/10.2136/sssaj2018.12.0472

Lorenz, M. O. (1905). Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association*, *9*(70), 209–219. https://doi.org/10.2307/2276207

Ma, Y., & He, H. (Eds.). (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st edition). Wiley-IEEE Press.

Mander, Ü. (n.d.). *Forest canopy mitigates soil N2O emission during hot moments | npj Climate and Atmospheric Science*. Retrieved February 17, 2024, from https://www.nature.com/articles/s41612-021-00194-7

Mare, D. S., Moreira, F., & Rossi, R. (2017). Nonstationary Z-Score measures. *European Journal of Operational Research*, *260*(1), 348–358. https://doi.org/10.1016/j.ejor.2016.12.001

Martiny, A. C., Talarmin, A., Mouginot, C., Lee, J. A., Huang, J. S., Gellene, A. G., & Caron, D. A. (2016). Biogeochemical interactions control a temporal succession in the elemental composition of marine communities. *Limnology and Oceanography*, *61*(2), 531–542. https://doi.org/10.1002/lno.10233

McClain, M. E., Boyer, E. W., Dent, C. L., Gergel, S. E., Grimm, N. B., Groffman, P. M., Hart, S. C., Harvey, J. W., Johnston, C. A., Mayorga, E., McDowell, W. H., & Pinay, G. (2003). Biogeochemical Hot Spots and Hot Moments at the Interface of Terrestrial and Aquatic Ecosystems. *Ecosystems*, *6*(4), 301–312. https://doi.org/10.1007/s10021-003-0161-9

Mcleod, E., Chmura, G. L., Bouillon, S., Salm, R., Björk, M., Duarte, C. M., Lovelock, C. E., Schlesinger, W. H., & Silliman, B. R. (2011). A blueprint for blue

carbon: Toward an improved understanding of the role of vegetated coastal habitats in sequestering CO2. *Frontiers in Ecology and the Environment*, *9*(10), 552–560. https://doi.org/10.1890/110004

McNicol, G., Fluet-Chouinard, E., Ouyang, Z., Knox, S., Zhang, Z., Aalto, T., Bansal, S., Chang, K.-Y., Chen, M., Delwiche, K., Feron, S., Goeckede, M., Liu, J., Malhotra, A., Melton, J. R., Riley, W., Vargas, R., Yuan, K., Ying, Q., … Jackson, R. B. (2023). Upscaling Wetland Methane Emissions From the FLUXNET-CH4 Eddy Covariance Network (UpCH4 v1.0): Model Development, Network Assessment, and Budget Comparison. *AGU Advances*, *4*(5), e2023AV000956. https://doi.org/10.1029/2023AV000956

Mentch, L., & Hooker, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research*, *17*(26), 1–41.

Middelburg, J. J., Nieuwenhuize, J., Iversen, N., Høgh, N., de Wilde, H., Helder, W., Seifert, R., & Christof, O. (2002). Methane distribution in European tidal estuaries. *Biogeochemistry*, *59*(1), 95–119. https://doi.org/10.1023/A:1015515130419

Mitsch, W. J., Bernal, B., Nahlik, A. M., Mander, Ü., Zhang, L., Anderson, C. J., Jørgensen, S. E., & Brix, H. (2013). Wetlands, carbon, and climate change. *Landscape Ecology*, *28*(4), 583–597. https://doi.org/10.1007/s10980-012-9758-8

Molodovskaya, M., Singurindy, O., Richards, B. K., Warland, J., Johnson, M. S., & Steenhuis, T. S. (2012). Temporal Variability of Nitrous Oxide from Fertilized Croplands: Hot Moment Analysis. *Soil Science Society of America Journal*, *76*(5), 1728–1740. https://doi.org/10.2136/sssaj2012.0039

Moncrieff, J., Clement, R., Finnigan, J., & Meyers, T. (2005). Averaging, Detrending, and Filtering of Eddy Covariance Time Series. In X. Lee, W. Massman, & B. Law (Eds.), *Handbook of Micrometeorology: A Guide for Surface Flux Measurement and Analysis* (pp. 7–31). Springer Netherlands. https://doi.org/10.1007/1-4020-2265-4_2

Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. *Comptes Rendus. Mécanique*, *347*(11), 845–855. https://doi.org/10.1016/j.crme.2019.11.009

Moore, T. R., & Knowles, R. (1989). THE INFLUENCE OF WATER TABLE LEVELS ON METHANE AND CARBON DIOXIDE EMISSIONS FROM PEATLAND SOILS. *Canadian Journal of Soil Science*, *69*(1), 33–38. https://doi.org/10.4141/cjss89-004

Morin, T. H. (2019). Advances in the Eddy Covariance Approach to CH4 Monitoring Over Two and a Half Decades. *Journal of Geophysical Research: Biogeosciences*, *124*(3), 453–460. https://doi.org/10.1029/2018JG004796

Nahm, F. S. (2022). Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36. https://doi.org/10.4097/kja.21209

Neubauer, S. C., & Megonigal, J. P. (2015). Moving Beyond Global Warming Potentials to Quantify the Climatic Role of Ecosystems. *Ecosystems*, *18*(6), 1000–1013. https://doi.org/10.1007/s10021-015-9879-4

Novick, K. A., Biederman, J. A., Desai, A. R., Litvak, M. E., Moore, D. J. P., Scott, R. L., & Torn, M. S. (2018). The AmeriFlux network: A coalition of the willing. *Agricultural and Forest Meteorology*, *249*, 444–456. https://doi.org/10.1016/j.agrformet.2017.10.009

Obregon, D., Mafa-Attoye, T. G., Baskerville, M., Mitter, E. K., de Souza, L. F., Oelbermann, M., Thevathasan, N. V., Tsai, S. M., & Dunfield, K. E. (2023). Functionality of methane cycling microbiome during methane flux hot moments from riparian buffer systems. *Science of The Total Environment*, *870*, 161921. https://doi.org/10.1016/j.scitotenv.2023.161921

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 154–168). Springer. https://doi.org/10.1007/978-3-642-31537-4_13

Pan, X.-Y., Wang, G.-X., Yang, H.-M., & Wei, X.-P. (2003). Effect of water deficits on within-plot variability in growth and grain yield of spring wheat in northwest China. *Field Crops Research*, *80*(3), 195–205. https://doi.org/10.1016/S0378-4290(02)00175-2

Panigrahi, R., & Borah, S. (2019). Dual-stage intrusion detection for class imbalance scenarios. *Computer Fraud & Security*, *2019*(12), 12–19. https://doi.org/10.1016/S1361-3723(19)30128-9

Parr. (n.d.). *Beware Default Random Forest Importances*. Retrieved March 8, 2024, from http://explained.ai/decision-tree-viz/index.html

Peacock, M., Gauci, V., Baird, A. J., Burden, A., Chapman, P. J., Cumming, A., Evans, J. G., Grayson, R. P., Holden, J., Kaduk, J., Morrison, R., Page, S., Pan, G., Ridley, L. M., Williamson, J., Worrall, F., & Evans, C. D. (2019). The full carbon balance of a rewetted cropland fen and a conservation-managed fen. *Agriculture, Ecosystems & Environment*, *269*, 1–12. https://doi.org/10.1016/j.agee.2018.09.020

Peltola, O., Vesala, T., Gao, Y., Räty, O., Alekseychik, P., Aurela, M., Chojnicki, B., Desai, A. R., Dolman, A. J., Euskirchen, E. S., Friborg, T., Göckede, M., Helbig, M., Humphreys, E., Jackson, R. B., Jocher, G., Joos, F., Klatt, J., Knox, S. H., … Aalto, T. (2019). Monthly gridded data product of northern wetland methane emissions based on upscaling eddy covariance observations. *Earth System Science Data*, *11*(3), 1263–1289. https://doi.org/10.5194/essd-11-1263-2019

Philippe, M. T., & Karume, K. (2019). Assessing Forest Cover Change and Deforestation Hot-Spots in the North Kivu Province, DR-Congo Using Remote Sensing and GIS. *American Journal of Geographic Information System*, *8*(2), 39–54.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, *9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, *20*(53), 1–32.

Ramachandran, K. M., & Tsokos, C. P. (2021). Chapter 5—Statistical estimation. In K. M. Ramachandran & C. P. Tsokos (Eds.), *Mathematical Statistics with Applications in R (Third Edition)* (pp. 179–251). Academic Press. https://doi.org/10.1016/B978-0-12-817815-7.00005-1

Reid, M. C., Tripathee, R., Schäfer, K. V. R., & Jaffé, P. R. (2013). Tidal marsh methane dynamics: Difference in seasonal lags in emissions driven by storage in vegetated versus unvegetated sediments. *Journal of Geophysical Research: Biogeosciences*, *118*(4), 1802–1813. https://doi.org/10.1002/2013JG002438

Rey-Sanchez, C., Arias-Ortiz, A., Kasak, K., Chu, H., Szutu, D., Verfaillie, J., & Baldocchi, D. (2022). Detecting Hot Spots of Methane Flux Using Footprint-Weighted Flux Maps. *Journal of Geophysical Research. Biogeosciences*, *127*(8), e2022JG006977. https://doi.org/10.1029/2022jg006977

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, *47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Rosentreter, J. A., Borges, A. V., Deemer, B. R., Holgerson, M. A., Liu, S., Song, C., Melack, J., Raymond, P. A., Duarte, C. M., Allen, G. H., Olefeldt, D., Poulter, B., Battin, T. I., & Eyre, B. D. (2021). Half of global methane emissions come from highly variable aquatic ecosystem sources. *Nature Geoscience*, *14*(4), 225–230. https://doi.org/10.1038/s41561-021-00715-2

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Saha, D., Kemanian, A. R., Montes, F., Gall, H., Adler, P. R., & Rau, B. M. (2018). Lorenz Curve and Gini Coefficient Reveal Hot Spots and Hot Moments for Nitrous Oxide Emissions. *Journal of Geophysical Research: Biogeosciences*, *123*(1), 193–206. https://doi.org/10.1002/2017JG004041

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A., & Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, *85*, 502–508. https://doi.org/10.1016/j.ecolind.2017.10.030

Saunois, M., Bousquet, P., Poulter, B., Peregon, A., Ciais, P., Canadell, J. G.,

Dlugokencky, E. J., Etiope, G., Bastviken, D., Houweling, S., Janssens-Maenhout, G., Tubiello, F. N., Castaldi, S., Jackson, R. B., Alexe, M., Arora, V. K., Beerling, D. J., Bergamaschi, P., Blake, D. R., … Zhu, Q. (2017). Variability and quasi-decadal changes in the methane budget over the period 2000–2012. *Atmospheric Chemistry and Physics*, *17*(18), 11135–11161. https://doi.org/10.5194/acp-17-11135-2017

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., Ciais, P., Arora, V. K., Bastviken, D., Bergamaschi, P., Blake, D. R., Brailsford, G., Bruhwiler, L., Carlson, K. M., Carrol, M., … Zhuang, Q. (2020). The Global Methane Budget 2000–2017. *Earth System Science Data*, *12*(3), 1561–1623. https://doi.org/10.5194/essd-12-1561-2020

Savage, K., Phillips, R., & Davidson, E. (2014). High temporal frequency measurements of greenhouse gas emissions from soils. *Biogeosciences*, *11*(10), 2709–2720. https://doi.org/10.5194/bg-11-2709-2014

Schmid, M., Ostrovsky, I., & McGinnis, D. F. (2017). Role of gas ebullition in the methane budget of a deep subtropical lake: What can we learn from process-based modeling? *Limnology and Oceanography*, *62*(6), 2674–2698. https://doi.org/10.1002/lno.10598

Schumacher, R. S., Hill, A. J., Klein, M., Nelson, J. A., Erickson, M. J., Trojniak, S. M., & Herman, G. R. (2021). From Random Forests to Flood Forecasts: A Research to Operations Success Story. *Bulletin of the American Meteorological Society*, *102*(9), E1742–E1755. https://doi.org/10.1175/BAMS-D-20-0186.1

Shahan, J., Chu, H., Windham-Myers, L., Matsumura, M., Carlin, J., Eichelmann, E., Stuart-Haentjens, E., Bergamaschi, B., Nakatsuka, K., Sturtevant, C., & Oikawa, P. (2022). Combining Eddy Covariance and Chamber Methods to Better Constrain CO2 and CH4 Fluxes Across a Heterogeneous Restored Tidal Wetland. *Journal of Geophysical Research: Biogeosciences*, *127*(9), e2022JG007112. https://doi.org/10.1029/2022JG007112

Sutton-Grier, A. E., Moore, A. K., Wiley, P. C., & Edwards, P. E. T. (2014). Incorporating ecosystem services into the implementation of existing U.S. natural resource management regulations: Operationalizing carbon sequestration and storage. *Marine Policy*, *43*, 246–253. https://doi.org/10.1016/j.marpol.2013.06.003

Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2010.5596486

Tian, W., Wang, H., Xiang, X., Loni, P. C., Qiu, X., Wang, R., Huang, X., & Tuovinen, O. H. (n.d.). Water table level controls methanogenic and methanotrophic communities and methane emissions in a Sphagnum-dominated peatland. *Microbiology Spectrum*, *11*(5), e01992-23. https://doi.org/10.1128/spectrum.01992-23

Tramontana, G., Ichii, K., Camps-Valls, G., Tomelleri, E., & Papale, D. (2015). Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data. *Remote Sensing of Environment*, *168*, 360–373. https://doi.org/10.1016/j.rse.2015.07.015

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., & Papale, D. (2016). Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, *13*(14), 4291–4313. https://doi.org/10.5194/bg-13-4291-2016

Tupek, B., Minkkinen, K., Pumpanen, J., Vesala, T., & Nikinmaa, E. (2015). CH_4 and N_2O dynamics in the boreal forest–mire ecotone. *Biogeosciences*, *12*(2), 281–297. https://doi.org/10.5194/bg-12-281-2015

Turner, J. C., Moorberg, C. J., Wong, A., Shea, K., Waldrop, M. P., Turetsky, M. R., & Neumann, R. B. (2020). Getting to the Root of Plant-Mediated Methane Emissions and Oxidation in a Thermokarst Bog. *Journal of Geophysical Research: Biogeosciences*, *125*(11), e2020JG005825. https://doi.org/10.1029/2020JG005825

Upstill-Goddard, R. C., Barnes, J., Frost, T., Punshon, S., & Owens, N. J. P. (2000). Methane in the southern North Sea: Low-salinity inputs, estuarine removal, and atmospheric flux. *Global Biogeochemical Cycles*, *14*(4), 1205–1217. https://doi.org/10.1029/1999GB001236

Valach, A. C., Kasak, K., Hemes, K. S., Anthony, T. L., Dronova, I., Taddeo, S., Silver, W. L., Szutu, D., Verfaillie, J., & Baldocchi, D. D. (2021). Productive wetlands restored for carbon sequestration quickly become net CO2 sinks with site-

level factors driving uptake variability. *PLoS ONE*, *16*(3), e0248398. https://doi.org/10.1371/journal.pone.0248398

Vegetation Affects Timing and Location of Wetland Methane Emissions—Bansal—2020—Journal of Geophysical Research: Biogeosciences—Wiley Online Library. (n.d.). Retrieved March 18, 2024, from https://agupubs-onlinelibrary-wiley-com.oca.ucsc.edu/doi/full/10.1029/2020JG005777

Vidon, P., Allan, C., Burns, D., Duval, T. P., Gurwick, N., Inamdar, S., Lowrance, R., Okay, J., Scott, D., & Sebestyen, S. (2010). Hot Spots and Hot Moments in Riparian Zones: Potential for Improved Water Quality Management1. *JAWRA Journal of the American Water Resources Association*, *46*(2), 278–298. https://doi.org/10.1111/j.1752-1688.2010.00420.x

Villa, J. A., Ju, Y., Stephen, T., Rey-Sanchez, C., Wrighton, K. C., & Bohrer, G. (2020). Plant-mediated methane transport in emergent and floating-leaved species of a temperate freshwater mineral-soil wetland. *Limnology and Oceanography*, *65*(7), 1635–1650. https://doi.org/10.1002/lno.11467

Villa, J. A., Ju, Y., Yazbeck, T., Waldo, S., Wrighton, K. C., & Bohrer, G. (2021). Ebullition dominates methane fluxes from the water surface across different ecohydrological patches in a temperate freshwater marsh at the end of the growing season. *Science of The Total Environment*, *767*, 144498. https://doi.org/10.1016/j.scitotenv.2020.144498

Waldo, S., Beaulieu, J. J., Barnett, W., Balz, D. A., Vanni, M. J., Williamson, T., & Walker, J. T. (2021). Temporal trends in methane emissions from a small eutrophic reservoir: The key role of a spring burst. *Biogeosciences*, *18*(19), 5291–5311. https://doi.org/10.5194/bg-18-5291-2021

Walter, J. A., Johnson, R. A., Atkins, J. W., Ortiz, D. A., & Wilkinson, G. M. (2023). Toward a Standardized Method for Quantifying Ecosystem Hot Spots and Hot Moments. *Ecosystems*, *26*(6), 1367–1378. https://doi.org/10.1007/s10021-023-00839-z

Weiner, J., & Solbrig, O. T. (1984). The meaning and measurement of size hierarchies in plant populations. *Oecologia*, *61*(3), 334–336. https://doi.org/10.1007/BF00379630

Wilson, S. R., Close, M. E., Abraham, P., Sarris, T. S., Banasiak, L., Stenger, R., & Hadfield, J. (2020). Achieving unbiased predictions of national-scale groundwater redox conditions via data oversampling and statistical learning. *Science of The Total Environment*, *705*, 135877. https://doi.org/10.1016/j.scitotenv.2019.135877

Woodrow, R. L., White, S. A., Sanders, C. J., Holloway, C. J., Wadnerkar, P. D., Conrad, S. R., Tucker, J. P., Davis, K. L., & Santos, I. R. (2022). Nitrous oxide hot moments and cold spots in a subtropical estuary: Floods and mangroves. *Estuarine, Coastal and Shelf Science*, *264*, 107656. https://doi.org/10.1016/j.ecss.2021.107656

Wu, X., Liu, H., Li, X., Ciais, P., Babst, F., Guo, W., Zhang, C., Magliulo, V., Pavelka, M., Liu, S., Huang, Y., Wang, P., Shi, C., & Ma, Y. (2018). Differentiating drought legacy effects on vegetation growth over the temperate Northern Hemisphere. *Global Change Biology*, *24*(1), 504–516. https://doi.org/10.1111/gcb.13920

Zhang, J., Chen, L., & Abid, F. (2019). Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method. *Journal of Healthcare Engineering*, *2019*, e7294582. https://doi.org/10.1155/2019/7294582