



## Research article

## Dispersion based recurrent neural network model for methane monitoring in Albertan tailings ponds

Esha Saha <sup>a</sup>, Oscar Wang <sup>b</sup>, Amit K. Chakraborty <sup>a</sup>, Pablo Venegas Garcia <sup>a</sup>,  
Russell Milne <sup>a</sup>, Hao Wang <sup>a</sup>

<sup>a</sup> Interdisciplinary Lab for Mathematical Ecology and Epidemiology (ILMEE) & Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G 2R3, Canada

<sup>b</sup> The Department of Mathematics and Computer Science, University of Richmond, Richmond, Virginia, 23173, United States

## ARTICLE INFO

Dataset link: <https://github.com/esha-saha/champ>

## Keywords:

Methane emissions

Oil sands tailing ponds

Atmospheric dispersion model

Neural networks

## ABSTRACT

Bitumen extraction for the production of synthetic crude oil in Canada's Athabasca Oil Sands industry has recently come under spotlight for being a significant source of greenhouse gas emissions. A major cause of concern is methane, a greenhouse gas produced by the anaerobic biodegradation of hydrocarbon in oil sands residues, or tailing, stored in settle basins commonly known as oil sands tailing ponds. In this work, we build a data-driven modeling framework to determine the methane emitting potential of these tailing ponds and have future methane projections using a Dispersion based Recurrent Neural Network (DIRNN). We show that our method can predict both methane emissions and concentrations by considering the transport of methane emissions in air, thereby outperforming existing other deep learning approaches. Using a reverse dispersion modeling approach, we use our trained model to identify active ponds and estimate about 56,303 tonnes of methane (1.5 million tonnes of carbon dioxide equivalent) emissions from the Athabasca oil sands tailings. Our results are consistent with previously reported emission estimates from various studies, and indicate at least three times underestimation in official reports.

## 1. Introduction

Anthropogenic sources of greenhouse gases (GHGs) are the major drivers of climate change with methane (CH<sub>4</sub>) having the second largest share of emissions in the atmosphere after carbon dioxide (CO<sub>2</sub>). Although the comparative impact of CH<sub>4</sub> is 28 times greater than CO<sub>2</sub> over a 100-year period (US EPA, 2024), it has a shorter lifespan of 12 years, which makes CH<sub>4</sub> mitigation policies a cost-effective short-term approach to combat global warming (Flannigan et al., 2009; Sysoeva et al., 2025). Oil sands activities contribute significantly to GHG emissions and in particular are regarded as sources of pollution (Schindler, 2014; Liggio et al., 2016; Yu and Zahidi, 2023). The mining and extraction of oil sands is directly associated with deforestation and release of sulfur oxides, nitrogen oxides, hydrocarbons, and fine particulate matter, etc. The oil sand tailing ponds (OSTPs) contain toxic industrial wastes that can leak into fresh water sources affecting aquatic ecosystems. Moreover, OSTPs emit significant quantities of CH<sub>4</sub> from toxin degradation by anaerobic bacteria (Siddique et al., 2007; Michel et al., 2024) leading to elevated levels of CH<sub>4</sub> in the air. Frequent onsite data collection for a detailed understanding of OSTPs

is often infeasible due to expensive measurement technology, toxic air quality or hard to obtain permit requirements (Sysoeva et al., 2025).

Prior OSTP related methane modeling efforts involve controlled laboratory experiments, whose results are used to build mechanistic models (MM) of CH<sub>4</sub> emissions (total mass of CH<sub>4</sub> released in atmosphere from source in a time interval) from toxin/hydrocarbon degradation by bacteria in OSTPs (Kong et al., 2019; Venegas Garcia, 2024). Since experimental limitations lead to exclusion of many relevant parameters such as temperature, pressure, wind speed, etc., MMs cannot be used to accurately assess methane concentration levels (the amount of CH<sub>4</sub> in the air at a given place and time). Infact, the relationship between emissions and concentrations can be better described by atmospheric dispersion models (ADMs) (Stockie, 2011; Das et al., 1998; Mohan and Siddiqui, 2002). However, simple ADMs like the Gaussian plume or puff models (Stockie, 2011; Mikkelsen et al., 1987) are unrealistic due to multiple modeling assumptions, while the realistic formulations are highly complex and non-linear, and may require high computational resources to be solved numerically. For example, refer to the model formulation of AERMOD (Cimorelli et al.,

\* Corresponding authors.

E-mail addresses: [esaha1@ualberta.ca](mailto:esaha1@ualberta.ca) (E. Saha), [hao8@ualberta.ca](mailto:hao8@ualberta.ca) (H. Wang).

2005), a well-known software developed by Environmental Protection Agency (EPA). One needs an understanding and data availability of multiple parameters woven through 30+ equations in order to solve the model. Other similar softwares include CALPUFF, FLEXPART, Wind-Trax, etc (Bakels et al., 2024; Bonifacio et al., 2013; Tagliaferri et al., 2022).

Data-driven techniques, on the other hand, offer a cost-effective and powerful alternative to classical modeling in the field of Environmental Sciences, both from the perspective of modeling complex phenomenon as well from the socio-economic modeling of policy gains. From a socio-economic and policy standpoint, data-driven approaches can enable low-cost, high-resolution tracking of industrial methane emissions, essential for stakeholders such as regulators, environmental agencies, and oil sands operators. Moreover, machine learning techniques are increasingly being applied to stock markets and energy pricing models (Kumar et al., 2025; Basher and Sadorsky, 2025; Jin and Xu, 2024; Cheng et al., 2025) including coal and carbon pricing models. These models help policymakers and analysts simulate the effects of emissions-based taxation under various economic conditions by capturing the non-linear dependencies between fuel demand, carbon markets, and policy shifts (Islam et al., 2025; Jin and Xu, 2024a; D'Orazio and Pham, 2025; Jin and Xu, 2025). These applications not only inform government strategy but also assist energy firms in optimizing operations.

For applications in modeling environmental phenomenon, some well-known machine learning architectures used for analyzing and predicting various atmospheric gases include Random Forests (RFs), neural networks (NN), recurrent neural networks (RNNs), long short-term memory (LSTM), bidirectional LSTM, stacked LSTM, and gated recurrent unit (GRU) (Xie et al., 2019; Tong et al., 2019; Luo et al., 2023; Hu et al., 2021; Hamrani et al., 2020; Meng et al., 2022; Hou et al., 2022). Other data-driven methods use satellite images as an effective way to build a dataset for the purpose of training classifiers to identify active OSTPs so that only high risk ponds may be closely monitored (Yu and Zahidi, 2023; Psomouli et al., 2023; Qu et al., 2024). Some recent works have used satellite data to detect CH<sub>4</sub> emissions across Canada (Yazdinejad et al., 2025; Zambrano-Luna et al., 2025). However, data collected by satellites is expensive, hard to process and sparse making them less useful for models where frequent data collection is the key to proper model training. Machine learning models can also be trained for effective risk prediction of OSTPs as an alternative to standard monitoring systems, which are often expensive and have poor lightning protection abilities (Yang et al., 2020). While most existing models (both mechanistic and data-driven) can perform well on estimating emissions from tailing ponds or predicting concentrations, an understanding of how atmospheric concentrations of CH<sub>4</sub> gets affected by emission sources are limited (Koushafar et al., 2023).

In order to fill this gap and track emission sources, we propose a novel Dispersion based Recurrent Neural Network (DIRNN) framework that can successfully use the dynamics of CH<sub>4</sub> emissions from OSTPs to predict concentrations in the atmosphere by preserving the physics of CH<sub>4</sub> dispersion in air. With a unique hybrid approach yet to be used widely in environmental modeling, given a dataset consisting of atmospheric variables (collected from weather monitoring stations around an OSTP) and hydrocarbon degradation (simulated from solving MMs) in an OSTP, our dispersion informed trained model can successfully predict CH<sub>4</sub> emissions and concentration levels near the OSTPs, outperforming other deep learning models. Instead of treating these two quantities separately, the proposed model enforces constraints based on atmospheric dispersion (advection–diffusion dynamics) to tie together emission and concentration. Model training using constrained optimization with penalty ensures balance between data fitting and physical consistency. Further, we show that our trained model is capable of giving emission estimates for various OSTPs, irrespective of whether they were included in the training dataset or not, as long as they are in a close proximity to each other. This is made possible due to the reverse dispersion penalty from physical equations that

enables daily emission estimation from multiple directions around weather stations, without direct sampling of each site. The novelty of our proposed model lies in various aspect of the framework which include (but are not limited to) model design (physics-constrained neural network using atmospheric transport), emission inference (inverse modeling of CH<sub>4</sub> sources from ambient data and wind direction), source attribution (directional decomposition of emissions), joint modeling (emissions and concentrations estimated simultaneously) and discovery of underestimated emissions, including from inactive ponds.

The paper is arranged as follows. We discuss the region of interest, data collection techniques and processing in Section 2, followed by the model framework in Section 3. The results are discussed in Section 4 with the discussion and conclusions in Sections 5 and 6, respectively.

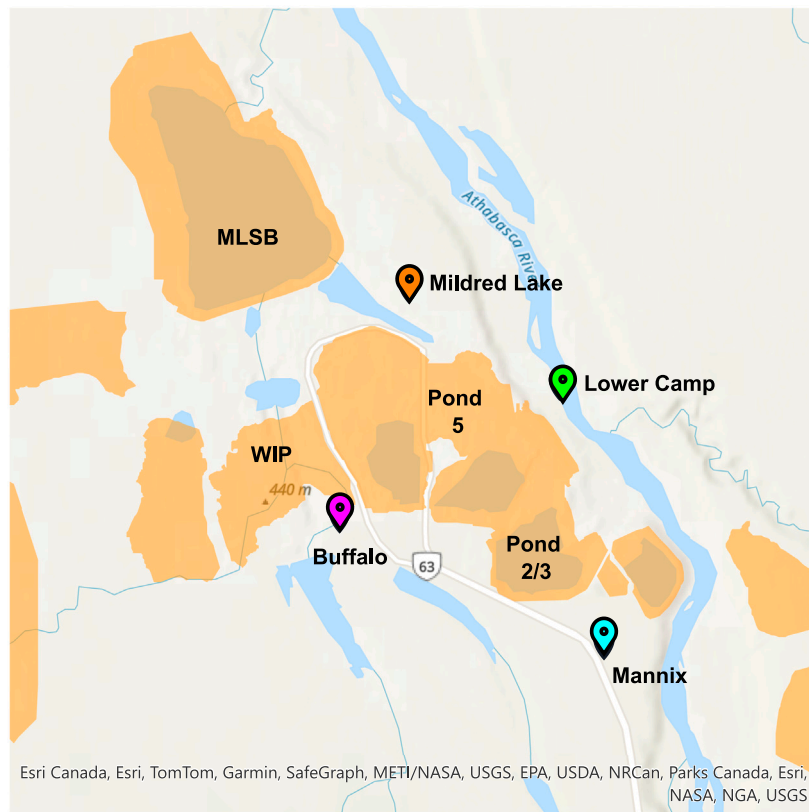
## 2. Study area, data collection and preprocessing

Oil sands tailings are the by-products generated after separating bitumen from oil sands. These tailings are stored in large engineered reservoirs known as oil sands tailings ponds (OSTPs) (Foght et al., 2017). They are composed of a mixture of sand, silt, clay fines, process water, and small amounts of unrecovered hydrocarbons from the extraction process (Gosselin et al., 2010). When first deposited into an OSTP, fresh tailings are mostly water (~ 85%), containing about 8% mineral fines and less than 1% unrecovered hydrocarbons. Over time, the fine particles gradually consolidate, eventually forming dense layers called mature fine tailings (MFT) with more than 30% solids near the pond bottom (Foght et al., 2017). Tailings temperature varies with depth, ranging from roughly 12 °C at 6 m (m) to about 22 °C at 30 m below the mudline (Penner and Foght, 2010; Ramos-Padrón et al., 2011). The process water is typically alkaline, with a pH of about 8.5. Chemically, tailings contain residual hydrocarbons and soluble electron acceptors such as sulfate and iron; in some cases, gypsum is added to accelerate consolidation. Direct data collection from OSTPs is challenging due to hard to obtain permit requirements.

Given a dataset consisting of atmospheric variables (collected from weather monitoring stations around an OSTP) and hydrocarbon degradation (simulated from solving the MMs built using laboratory controlled experiments) in an OSTP, we are interested in building a machine learning model that can predict both methane concentration and emission simultaneously. Thus, the input to the proposed framework includes various parameters that directly or indirectly affect atmospheric methane concentrations. The model considers three types of input data: (i)  $x_{dil}$  denoting the degradation of hydrocarbons in OSTPs and obtained from solving MMs in literature; (ii)  $x_{atm}$  representing atmospheric parameters such as ambient temperature, wind speed, wind direction, solar activity, etc; and (iii) time vector  $t$ . These three inputs  $x_{dil}$ ,  $x_{atm}$ , and  $t$  together form the input  $x$  and are used to define the model and its corresponding minimization problem. In this section, we discuss the study area (Section 2.1) and the technique of dataset building. The meteorological data collection from weather monitoring stations is discussed in Section 2.2 and simulated data from experimentally validated MM (which estimates methane emissions and hydrocarbon degradation in active OSTPs) is discussed in Section 2.3.

### 2.1. Study area

Our region of interest is located in the industrial area around Syncrude and Suncor Base Plants in the Athabasca Oil Sands deposits. The region contains multiple weather monitoring stations (some located near the oil sands mining areas) under the Wood Buffalo Environmental Association (WBEA) (Wood Buffalo Environmental Association, 2024) that measures hourly data of the ambient air quality and meteorological parameters. Overall elevation of the WBEA region is about 200–300 m above sealevel. Lower Camp is located by the Athabasca River Valley at about 115 m south of the Syncrude pump house and 238 m above sealevel. It has an active OSTP ‘Pond 2/3’ approximately at a distance



**Fig. 1.** Region of Wood Buffalo with all the weather monitoring stations of interest and main OSTPs and/or EPLs. Selected stations: Mannix ('Pond 2/3' approximately 1.4 km northwest of the station); Lower Camp ('Pond 2/3' approximately 3.5 kms southwest and 'Pond 5' about 1.4 kms west of the station, respectively); Mildred Lake ('Mildred Lake Settling Basin'/MLSB approximately 1.7 kms northwest and 'Pond 5' approximately 2.1 kms southeast of the station, respectively); Buffalo ('West-In-Pit'/WIP at 0.8 kms northwest of the station). The orange polygons depict the oil sands tailings ponds. The blue water inside the orange polygons depict the main water body. However, since tailings also include a semi-solid mixture of sand, silt, clay fines, and unrecovered industrial chemicals around the water-filled area, the overall area of tailings may differ and is highlighted separately in orange.

of 3.5 kms southwest of the station and an abandoned OSTP 'Pond 5' approximately at 1.4 kms, both owned by Suncor. Located at 332 m above sea level, Mannix station is less than 5 km from the Suncor base plant whose land use segregation is between 0–180 degrees of the station. There are no airflow restrictions and an active OSTP 'Pond 2/3' owned by Suncor is located approximately 1.4 kms northwest of the station. Mildred Lake is Located within 400 m of the Syncrude airstrip, the station sits at 314 m above sea level and within 5 km from the Syncrude base plant on West. There is partial restriction of airflow in North, South and West of the station by buildings and/or trees which lie within 40–160 m. It has been measuring methane from December 2019 onwards. An active OSTP 'Mildred Lake Settling Basin'/MLSB is approximately 1.7 kms northwest of the station, owned by Syncrude, and inactive 'Pond 5' is approximately 2.1 kms owned by Suncor. Buffalo station sits at 315 m above sea level and less than 5 km from the Syncrude base plant and 0.8 km from an OSTP. There is no restriction of airflow. The land use segregation reports oil sands plant in 0–90 degrees and 271–360 degrees of the station. It is nearest pond, 'West-In-Pit'/WIP is presently converted to an EPL and sits at a distance of 0.8 kms northwest of the station. The study area is given in Fig. 1.

Note that out of the four selected stations, we build datasets using the stations that are associated with 'active' OSTPs only, because source estimations using MMs are based on experiments modeled after 'active' OSTPs. We classify an OSTP to be 'active' if it there is continuous inflow of diluents (industrial residues). Based on available data and information on OSTPs, we select stations Mannix with Pond 2/3, Lower Camp with Pond 2/3 and Mildred Lake with MLSB for building the dataset.

## 2.2. Dataset creation from on-field data

For  $d > 1$  and number of days  $k$ , suppose the input–output pairs are denoted by  $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^k$ , with  $\mathbf{x} = [\mathbf{x}_{atm}, \mathbf{x}_{dil}] \in \mathbb{R}^d$  and  $\mathbf{y} = [y_n^{conc}, y_n^{emm}] \in \mathbb{R}^2$ . The vector  $\mathbf{x}_n$  consists of various input variables sampled across  $k$  timesteps (days) and the two-dimensional vector  $\mathbf{y}_n$  consists of the  $\text{CH}_4$  emissions, and concentrations. For each  $k$ , the input vector  $\mathbf{x}_n$  is built using data from two sources: atmospheric variables  $\mathbf{x}_{atm}$  and industrial/chemical component variables  $\mathbf{x}_{dil}$ . Dataset for atmospheric parameters  $\mathbf{x}_{atm}$  and methane concentrations  $y_n^{conc}$  is built from data collected by the Wood Buffalo Environmental Association (WBEA) (Wood Buffalo Environmental Association, 2024). The training dataset is built only for stations near an active OSTP (MLSB and Pond 2/3 as reported in Burkus et al. (2014)) (as the mechanistic models can only mimic the kinetics of an active tailings pond). We pick the stations in a close proximity to these active OSTPs that have no other methane sources (for example, wetlands) between them.

For each station, we build the dataset by filtering observations with wind direction ranges based on location of active OSTPs i.e., Mannix (300–320 degrees), Lower Camp (160–180 degrees), and Mildred Lake (300–340 degrees). The data from atmospheric variables denoted by  $\mathbf{x}_{atm}$  includes temperature, pressure, relative humidity, solar radiation, etc and is collected hourly by weather monitoring stations under consideration. Similarly, the WBEA dataset reporting  $\text{CH}_4$  concentrations is used to build one of the output dataset samples,  $y_n^{conc}$ . The map of the weather stations and OSTPs is given in Fig. 1. All variables used in the input  $\mathbf{x}_{atm}$  is summarized in Table C.6.

### 2.3. Data from experimentally validated model(s)

To build dataset for  $\mathbf{x}_{dil}$  and source emissions  $y_n^{emm}$ , we solve MMs describing the properties and dynamics of methane production in OSTPs, with appropriate parameters and initial conditions (Kong et al., 2019; Siddique et al., 2008; Venegas Garcia, 2024). These experimentally validated models are developed by considering the most labile hydrocarbons present in the diluents/solvents used by each of the oil sands companies. The models are generally represented by a dynamical system whose general form is described in Eq. (1). The system describes the degradation dynamics of each of the labile hydrocarbon by methanogenic bacteria. For each fixed  $i$  (the value of  $i$  depends on the number of labile hydrocarbons considered), the system of equations are given as

$$\begin{aligned} \frac{dC_i}{dt} &= f(C_i, t, y_1, \dots, y_k) \\ \frac{dy_j}{dt} &= g_j(C_i, t, y_j) \text{ for } j = 1, \dots, k \\ CH_4 &= h(C_i, y_1, \dots, y_k, \mu_i, t) \end{aligned} \quad (1)$$

where  $C_i$  denotes each of the hydrocarbons,  $y_j$  denotes other variables in consideration (for example, other nutrients, biomass of microbes, etc.) and  $\mu_i$  denotes the set of constants corresponding to methane production (for example, microbial efficacy, stoichiometric factor, etc.). The exact nature of the functions  $f, g$  and  $h$  along with the other parameters can be found in prior works on modeling methane emissions from OSTPs (Kong et al., 2019; Siddique et al., 2008; Venegas Garcia, 2024).

In order to solve the MM, we use the monthly ‘Flared and Wasted’ category of “diluents” from (Alberta Energy Regulator, 2023) as the total monthly inflow of hydrocarbons into the ponds. Based on prior works that have studied possible chemical composition of these diluents (Kong et al., 2019; Siddique et al., 2006), we split this diluent data into about 20 labile hydrocarbons. Note that we consider only a small fraction of the diluents based on the OSTP Fine Fluid Tailings (FFT) volume (20% for MLSB and 15% for Pond 2/3 i.e., only this much amount of total reported diluents is assumed to be tossed into the OSTPs) and divide it by the number of days per month to estimate daily inflow of diluents in the MM. These hydrocarbons are assumed to be subsequently used up by methanogenic bacteria leading to methane emissions. A constant daily inflow of diluents is assumed i.e. the total monthly diluent reported in Alberta Energy Regulator (2023) is divided equally by the number of days in the month (to obtain the initial conditions). The system is solved with a timestep of one day for a month. This is done for all the months from January 2020 to December 2023 in order to build the dataset. This approach is similar to the technique followed in Venegas Garcia (2024). The hydrocarbon degradation described by the MMs forms the dataset  $\mathbf{x}_{dil}$ . We experiment with  $\mathbf{x}_{dil}$  built based on two methanogenesis models (Venegas Garcia, 2024; Kong et al., 2019). The  $CH_4$  values obtained using these models form  $y_n^{emm}$ . Note that when generating the data using the model proposed in Kong et al. (2019), we identified a small number of generated daily values (about 1% of values) that were not biologically realistic. We considered these values to be numerical artifacts of simulation driven by the stiffness of the model, and hence replaced each one with the generated value for the day before it for the sake of simplicity and biological fidelity.

Once all types of data is obtained, they are averaged and interpolated to obtain sample points at the frequency of one data per day. For example, variables from  $\mathbf{x}_{atm}$  are collected hourly, so they are averaged over 24 h to obtain one sample per day. On the other hand, diluent data is reported monthly by companies and hence interpolation technique is used to fit a spline function and sample daily data points.

### 3. Problem formulation and framework

The goal of the proposed research is to train a parameterized model to track emissions from OSTPs using  $CH_4$  concentrations. We incorporate the dynamics from MMs of OSTPs and physical constraints from atmospheric dispersion models to train a machine learning framework over a give period of time. The idea is to model the interactions depicted in Fig. 2.

- Diluents directly affect emissions from OSTPs. For example, more diluents would directly lead to more hydrocarbon degradation and thus more  $CH_4$  emissions.
- $CH_4$  emissions and concentrations are directly related to each other. Increase in one of the quantities will automatically lead to an increase in the other and vice versa.
- Atmospheric parameters directly affect the concentration of  $CH_4$  but do not significantly affect the formation of  $CH_4$  in the OSTPs (for example, the wind direction or speed has nothing to do with how the microbes degrade hydrocarbons at the bottom layer of the OSTPs).
- Diluents indirectly affect the concentrations by increasing emissions which in turn increases the concentrations.

Thus, we combine all the variables (diluents and atmospheric data) and use it as an input to a machine learning model. The output is then combined with the data from atmospheric variables (only) and used as an input to another machine learning model which is defined using the atmospheric dispersion model equation(s). A detailed flowchart is given in Fig. 2.

To optimize the weights and biases of DIRNN, the dispersion constrained optimization problem (Xu and Darve, 2022; Saha et al., 2023a; Kashinath et al., 2021; Antonion et al., 2024) is modified to include information from MMs as well as real-time data affecting concentrations. For each  $i$ th observations in the set  $I_{obs}$ , suppose  $\mathbf{x}_i$  denotes the  $(d+1)$ -dimensional input vector and  $u_i$  denotes observed output. Then given a fixed function  $q: \mathbb{R}^{d+1} \rightarrow \mathbb{R}$  describing emission dynamics (from MMs), our modified constrained optimization problem aims to find a function  $u: \mathbb{R}^{d+1} \rightarrow \mathbb{R}$  by solving the problem,

$$\min_{\phi} \frac{1}{|I_{obs}|} \sum_{i \in I_{obs}} (u(\mathbf{x}_i) - u_i)^2 \text{ subject to } F(\phi(\mathbf{x}, u), u, q) = 0, \quad (2)$$

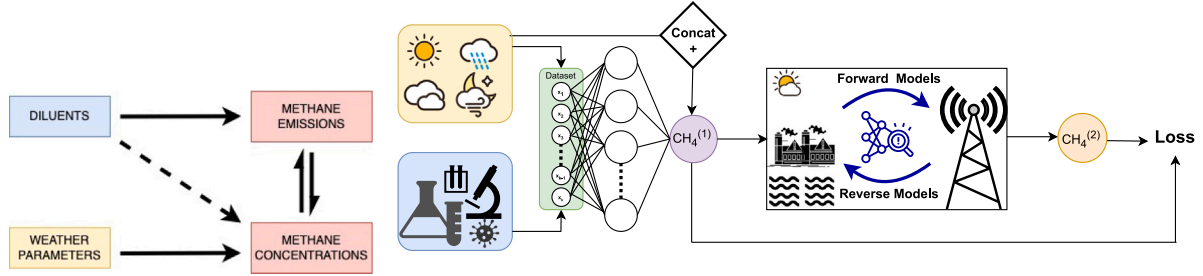
where  $F$  is the physical constraint with unknown function  $\phi$ . Since we want to learn  $u$  from given measurements and  $\phi$  is unknown, we parameterize them with  $u_{\hat{\theta}}(\mathbf{x})$  and  $\phi_{\hat{\theta}}(\mathbf{x}, u)$ , respectively where  $\hat{\theta}$  and  $\bar{\theta}$  denote the unknown parameters to be learned through training. Different function representation models such as sparse polynomial approximation (Schaeffer et al., 2018), random feature models (Saha et al., 2023b; Saha and Tran, 2023), or neural networks (McCulloch and Pitts, 1943) may be used. Converting Eq. (2) to an unconstrained optimization problem we get,

$$\min_{\hat{\theta}, \bar{\theta}} \frac{1}{|I_{obs}|} \sum_{i \in I_{obs}} \left[ (u_{\hat{\theta}}(\mathbf{x}_i) - u_i)^2 + \lambda (F(\phi_{\bar{\theta}}(\mathbf{x}_i, u_{\hat{\theta}}), u_{\hat{\theta}}, q(\mathbf{x}_i)))^2 \right], \quad (3)$$

where  $\lambda \in (0, \infty)$ . An advantage of using this method (also known as penalty method) is that it avoids solving the constraint  $F(\phi, u, q) = 0$  (Xu and Darve, 2022). However, the physical constraints may not be satisfied exactly i.e., theoretically,  $F(\phi, u, q) = 0$  will hold only when  $\lambda \rightarrow \infty$  (Xu and Darve, 2022). While optimizing the choice of  $\lambda$ , it is important to remember that a large value of  $\lambda$  places less weight on the objective function. Hence, a proper choice of  $\lambda$  is based on the desirable trade-off between fitting the observed value and satisfying the constraint. This technique has been adapted in numerous works involving learning of systems of PDEs/ODEs from data (Raissi et al., 2019). Our choice of the constraint  $F$  is derived from an atmospheric dispersion models called Gaussian Plume Model (GPM) (Stockie, 2011),

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{J} = q,$$





**Fig. 2.** Graphical representation of the proposed modeling framework. **Left:** Representation of how the different input variables interact. Solid lines depict a direct connection and the dashed lines represent an indirect affect. **Right:** A flowchart of DIRNN.

where  $u(\vec{x}, t)$  is the mass concentration,  $q(\vec{x}, t)$  is a source (or sink) and  $J$  is mass flux due to diffusion ( $J_D$ ) and advection ( $J_A$ ), and  $\vec{x}$  and  $t$  denote space and time respectively. Assuming negligible sinks, the function  $q$  defined from the MMs acts as the source term. Since we fix the coordinates of the source (OSTPs) and weather stations in space,  $u$  and  $q$  are independent of spatial coordinates. Thus, we are interested in exploring the relationship of these two functions ( $u$  and  $q$ ) to various input variables involving atmospheric parameters and/or hydrocarbon degradation.

Given the input dataset  $\mathbf{x} = [\mathbf{x}_{atm}, \mathbf{x}_{dil}, \mathbf{t}]^T \in \mathbb{R}^{d+1}$ , where  $\mathbf{x}_{dil}$  is built from simulated data from MMs and  $\mathbf{x}_{atm}$  is built from the real-time weather station measurements, the  $\text{CH}_4$  concentrations and emissions are trained using the parameterization below in Eq. (4) and (5). Output of the first network  $u_{\hat{\theta}}$  gives the predicted methane concentration. We then concatenate this output with atmospheric parameters  $\mathbf{x}_{atm}$  (since those are the only variables that can potentially affect diffusion/advection), and use it as an input to the second neural network that estimates  $\nabla J$ . The exact form of the outputs are given by

$$u_{\hat{\theta}} = \hat{\theta}_4 \sigma (\hat{\theta}_3 \sigma (\hat{\theta}_2 \sigma (\hat{\theta}_1 \mathbf{x}))) \quad (4)$$

$$q_{\hat{\theta}}(\mathbf{x}, t) = \text{Grad}_t (u_{\hat{\theta}}) + \hat{\theta}_3 \sigma (\hat{\theta}_2 \sigma (\hat{\theta}_1 [u_{\hat{\theta}}, \mathbf{x}_{atm}])) \quad (5)$$

where  $\sigma$  denotes the activation function,  $\text{Grad}_t(\cdot)$  denotes the partial time derivative of the output  $u_{\hat{\theta}}$  and  $q_{\hat{\theta}}(\mathbf{x}, t)$  denotes the learned parameterized source emissions. Alternatively, the function to define  $\nabla J$  can also be represented with a known basis such as a polynomial basis. This can be used when trying to decipher the important variables affecting the diffusion/advection terms. Eq. (5) thus modifies to

$$q_{\hat{\theta}}(\mathbf{x}, t) = \text{Grad}_t (u_{\hat{\theta}}) + \hat{\theta}_1 \mathcal{P}([u_{\hat{\theta}}, \mathbf{x}_{atm}]),$$

where  $\mathcal{P}([z_1, \dots, z_n]) = [1, z_1, \dots, z_n, z_1 z_2, \dots, z_{n-1} z_n, z_1^2, \dots, z_n^2]$ .

### 3.1. Inverse dispersion based framework (iDIRNN) for emission estimates

In order to estimate source emissions and identify active tailing, we modify the constraint to satisfy inverse dispersion models that aim to quantify emissions when concentrations are given. Given a source emitting gas at a continuous and unknown rate  $q$  kg/s, suppose the time-average gas concentration above background denoted by  $C - C_b$  (where  $C$  is the measured concentration and  $C_b$  is the background concentration) is measured at some point  $M$ . Then the emission using inverse dispersion modeling is given as

$$q = (C/Q)_{sim}^{-1} (C - C_b),$$

where  $(C/Q)_{sim}$  denotes the ratio of concentration at  $M$  to the source emission rate predicted by an atmospheric dispersion model. While the equation seems pretty straightforward, prediction of  $(C/Q)_{sim}$  is not trivial and often ill-conditioned. Different types of dispersion models (e.g., Gaussian plume, K-theory) make this calculation with different levels of sophistication. Realistically, models should take into account average wind and turbulence statistics of the atmosphere along with

possible dispersion of the source and how it relates to the concentrations. Various numerical methods as well as prior emission estimates may be needed to solve the problem (Vojta et al., 2022).

In iDIRNN, using inputs  $[\mathbf{x}_{dil}, \mathbf{x}_{atm}, \mathbf{t}]^T \in \mathbb{R}^{d+1}$  we still learn  $u_{\hat{\theta}}$  from fitting the model to observed concentrations as in Eq. (4), however the constraint is now based on finding source emissions  $q$  from the above equation. We parameterize the inverse of influence function  $(C/Q)_{sim}$  as a neural network so that the constraint in Eq. (5) becomes,

$$q_{\hat{\theta}}(\mathbf{x}, t) = \hat{\theta}_3 \sigma (\hat{\theta}_2 \sigma (\hat{\theta}_1 [u_{\hat{\theta}}, \mathbf{x}_{atm}])) \quad (6)$$

Note that in this formulation, the term  $C_b$  term is balanced out by the bias terms present in estimating emissions and concentrations. This offers multiple advantages: (1) the formulation lets the model automatically learn from given data; (2) it is useful for various applications where data and understanding of the dynamics is limited. We first train the model on our datasets. In order to get emission estimates, we replace  $u_{\hat{\theta}}$  in Eq. (4) with the concentration data measured by the weather station into the trained model. The emission estimates per day are then added up to get cumulative emissions over a given year.

### 3.2. Model architecture, training and validation analysis

#### Algorithm 1 Model training using proposed framework

**Require:** Input data  $\mathbf{x} = [\mathbf{x}_{dil}, \mathbf{x}_{atm}, \mathbf{t}]$ , observed data  $u$ , emission function  $q(\mathbf{x})$ , observations indices set  $I_{obs}$ , physical constraint  $F$ , model architectures for learning  $u_{\hat{\theta}}$  and  $\phi_{\hat{\theta}}$ , penalty parameter  $\lambda$ .

- 1: **for**  $j$  in epochs **do**
- 2:    $y = u_{\hat{\theta}}(\mathbf{x})$ ;  $z = \phi_{\hat{\theta}}(y, \mathbf{x}_{atm})$ .
- 3:   Update  $\Theta = [\hat{\Theta}, \hat{\Theta}]$  by minimizing the loss,

$$\frac{1}{|I_{obs}|} \|y - u\|_2^2 + \frac{\lambda}{|I_{obs}|} \|F(z, y, q)\|_2^2$$

- 4:   **if** Sparse Parameters == True **then**
- 5:     **for**  $k$  in size( $\hat{\Theta}$ ) **do**
- 6:       **if**  $|\hat{\Theta}_k| < 10^{-4}$  **then**
- 7:          $\hat{\Theta}_k = 0$
- 8:       **end if**
- 9:     **end for**
- 10:   **end if**
- 11: **end for**

**Output:** Concentrations  $y$  and emissions  $z$

The entire dataset contains 1096 daily measurements dated between January 1, 2020 to December 31, 2022. Dataset for the year 2023 is built separately for testing the trained model(s), consisting of 365 samples to predict emissions and concentrations from the trained model. The training dataset is standardized between 0 and 1 to avoid unnecessary bias of input features with a larger scale. The conversion between different units of measurement between emission and concentration

data are implicitly taken care of while model training. The model was trained on 80% of the samples from the dataset with up to 10,000 iterations. The training loss for all the stations fell below  $10^{-4}$ .

Given the size of our dataset, we use three layers with 500 and 200 neurons each to learn  $u_{\theta}$  and  $\phi_{\theta}$ , respectively. For optimizing the weights, we use a stochastic gradient descent (SGD) algorithm with learning rate optimized between  $10^{-2}$  and  $10^{-3}$ , and momentum 0.9. Since the data from weather monitoring stations are noisy, we use an  $\ell_2$  based weight decay parameter of  $10^{-3}$  in the SGD algorithm to avoid overfitting. For 10 random initializations, we train the iDIRNN model and calculate the estimated emissions using real concentration data for the years 2020 and 2021 (true emission data from companies is only available for these two years). The model whose cumulative  $\text{CH}_4$  estimations are closest to the true emissions as reported in official documents is picked for further analysis. All codes were written in Python 3.10.9 and can be found at <https://github.com/esha-saha/champ>. The pseudocode for model training is given in Algorithm 1. We extrapolate our results to obtain total emissions from all OSTPs using the sample mean of emissions computed for the selected OSTPs. Since official statistical data reports  $\text{CH}_4$  emissions in terms  $\text{CO}_2$  equivalent, we convert our estimations to similar units accordingly for comparison.

## 4. Results

We discuss the model simulation outcomes with respect to two objectives: (i) forecasting methane emissions and concentrations from tailing ponds jointly; (ii) identifying active OSTPs and their emissions levels to analyze their impacts with respect to the overall  $\text{CH}_4$  emissions in Canada.

### 4.1. Methane forecasts from tailing ponds

Given the mass of industrial waste in a given OSTP and the meteorological conditions around them, DIRNN is trained to predict  $\text{CH}_4$  emissions from the pond and its corresponding concentration at the closest weather monitoring station. Fig. 3 and Table 1 suggest that different representations have similar predictive power. Other representations such as a simple Recurrent Neural Network (RNN) without dispersion constraints, Long-Short-Term-Memory (LSTM) (as defined in Appendix A) performed poorly in comparison to our proposed model. The results demonstrate a seasonal and diluent-dependent relationship between methane concentrations and source emissions using a data-driven ADM. We found an increasing/decreasing trend of methane concentration and emissions based on input data consisting of atmospheric variables and hydrocarbon degradation data. A direct connection between emissions and concentrations is established through our model indicating a rise in atmospheric  $\text{CH}_4$  concentration levels right after increased emissions. Majority of our predictions fall within the 95% confidence bound, which was found to be narrower on training data as expected, and comparatively wider for the validation set as those predictions are based on unseen data. Since the concentration data is noisy, we also see noise/oscillations in fitting of the emission data, which is discussed in detail in Section 5.

As with any machine learning algorithm, the model training is affected by the quality of training data which is what creates a difference in performance depending on the station under consideration. For  $\text{CH}_4$  concentration data station Mannix has the highest variance of 0.12 with maximum and average  $\text{CH}_4$  concentrations being 4.56 ppm and 2.21 ppm respectively, followed by station Mildred Lake with a variance of 0.05, and maximum and average  $\text{CH}_4$  concentrations being 3.95 ppm and 2.1 ppm respectively. Station Lower Camp has lowest variance of 0.03 with maximum and average  $\text{CH}_4$  concentrations being 3.36 ppm and 2.001 ppm respectively. From the relative errors in Table 1, it can be seen that lower the variance, better the performance of the models. This happens as data-driven models tend to have larger biases in predicting the extreme values, especially in data-scarce training

**Table 1**

Average of relative training and validation errors for predicting concentrations and emissions from each station and its corresponding tailing pond. For each column (station), the two lowest errors are highlighted. DIRNN<sub>forward</sub> and DIRNN<sub>poly</sub> denotes the training framework that uses forward dispersion model to predict concentrations from emissions. iDIRNN refers to the framework using reverse dispersion model that learns the emissions given the concentrations.

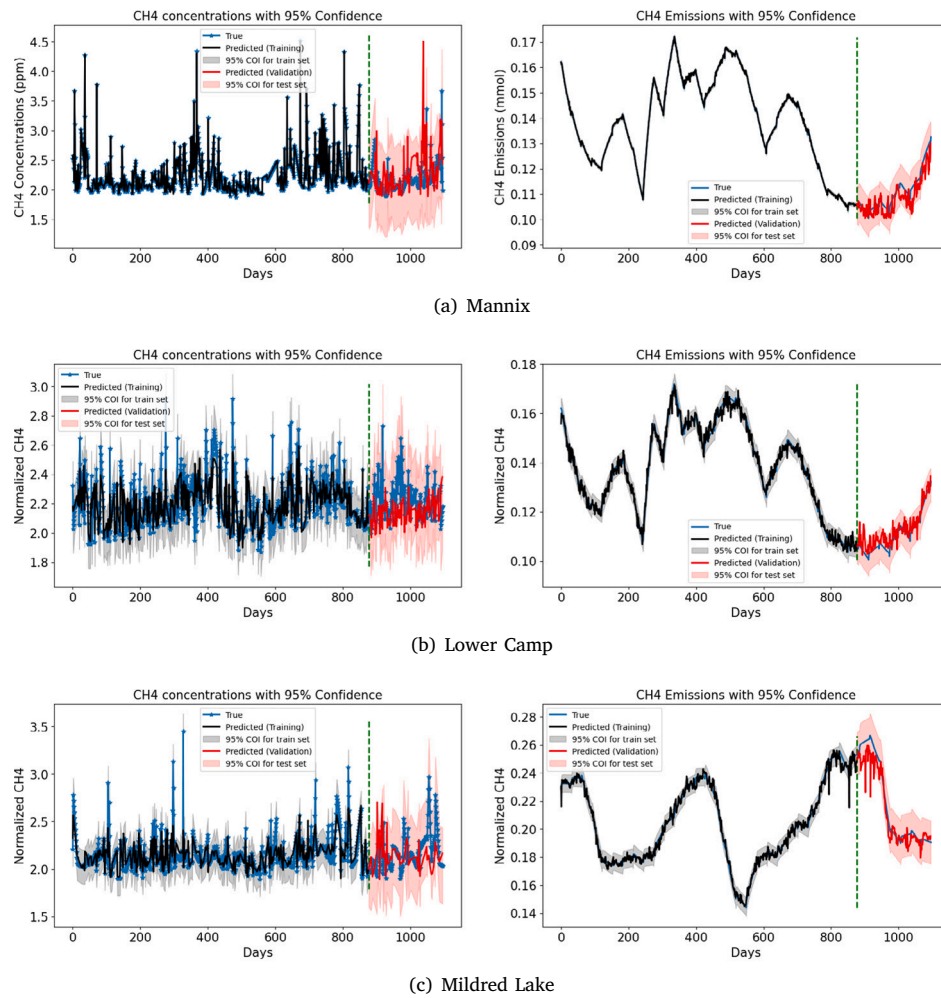
|                          | Concentration |               |               | Emission      |               |               |
|--------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                          | Mannix        | LC            | ML            | Mannix        | LC            | ML            |
| DIRNN <sub>forward</sub> | <b>0.5425</b> | <b>0.3101</b> | 0.5814        | <b>0.0992</b> | 0.1615        | <b>0.0517</b> |
| iDIRNN                   | 0.6566        | 0.4301        | 0.5702        | 0.1085        | <b>0.1369</b> | <b>0.0543</b> |
| DIRNN <sub>poly</sub>    | <b>0.4624</b> | <b>0.3140</b> | <b>0.4225</b> | 0.5380        | 0.1977        | 0.0942        |
| RNN                      | 0.6239        | 0.4314        | 2.1859        | <b>0.0661</b> | <b>0.0737</b> | 0.7127        |
| LSTM                     | 0.6454        | 0.3914        | <b>0.5550</b> | 0.2426        | 0.2568        | 0.0711        |

regimes such as in this scenario. However, as we use other physical constraints in our framework, the model can outperform other models even with datasets with higher levels of noise and variation.

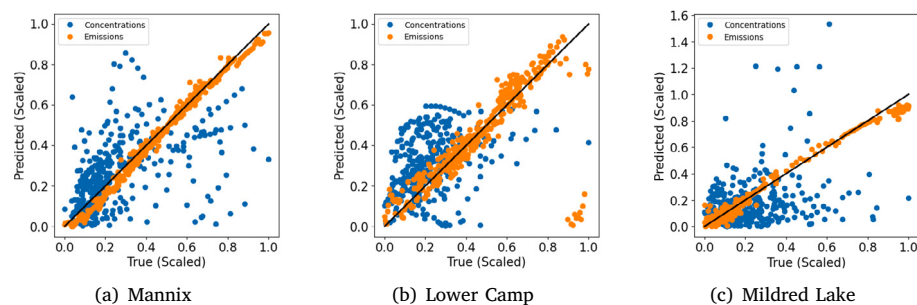
To test our model's long term  $\text{CH}_4$  concentration predictive capacity, we obtain predictions for the year 2023 and compare it with the observed data. Fig. 4 shows that the trained model accurately estimates emissions from OSTPs up to one year ahead. For predicting concentrations, the model can suggest future trends, with better data fitting when the true data is closer to the mean  $\text{CH}_4$  levels. However, the variation is larger (for higher recorded observation) since the real-time data is noisy and the model ignores extreme large values as outliers. Our results indicate that with current levels of oil sands activity and similar meteorological parameter readings (temperature, humidity, etc.), there will be no improvement in  $\text{CH}_4$  concentrations in the region.

### 4.2. Identification of $\text{CH}_4$ sources around weather stations

Due to the Athabasca region being located so further up in the North, a lot of the existing remote sensing methods for continuous monitoring of emissions are not directly applicable. The satellite data is often of poor quality (due to low sunlight, especially during winters and/or cloud cover). Other data measurement techniques (example, airplanes or drones) generally do not differentiate between multiple sources of emissions and are mostly expensive to carry out on a regular basis. Thus, to build a  $\text{CH}_4$  emission monitoring method, we train the proposed model using the reverse formulation of ADM to estimate daily emissions 360 degrees around a selected weather station. Once trained, we replace the input  $u_{\theta}$  in Eq. (6) with true  $\text{CH}_4$  concentration data from weather monitoring stations to obtain daily emission estimates from the tailing ponds. Daily predicted emissions are summed up to obtain cumulative emissions for each year between 2020 to 2023 and is given as a radial plot in Figure C.27. To compare emission levels over a period of three years, we plot the emissions for the year 2020 and 2023 in Fig. 5. The two major reported OSTPs (MLSB and Pond 2/3) and abandoned pond/EPLs (Pond 5 and WIP) are marked on the map. For weather monitoring stations in the vicinity of inactive ponds, we used a trained model from another appropriate station to get emission estimates. For example, since WIP lake (an inactive pond) near station Buffalo is owned by Syncrude, we can use a model trained with Syncrude's data to get emission estimates around it. Here we used the model trained for the Mildred Lake station. For each station we see the estimated emissions are highest from the direction of tailing ponds. Both Mannix and Lower Camp indicate that  $\text{CH}_4$  emissions from Pond 2/3 are more than 850 tonnes (t) per year every 20 degree interval. Each year, station Mildred Lake captures more than 4500 t of  $\text{CH}_4$  emissions from the direction of MLSB. The other inactive ponds (WIP and Pond 5) are often ignored as a significant sources of  $\text{CH}_4$  emissions as there is no inflow of diluents. However, using the dataset of station Buffalo, iDIRNN estimates more than 2000 t of emissions coming from the direction of WIP lake and about 1900 t of  $\text{CH}_4$  from Pond 5. These



**Fig. 3.** Results for predicting  $\text{CH}_4$  concentrations and emissions using a neural network based representation of the dispersion/advection terms in ADM. The green dashed line indicates the data split between training and validation set. For concentrations, the model accurately captures the seasonal nature of concentrations as well as the emissions based on given input data. The model predictions always fall within the 95% confidence of interval.

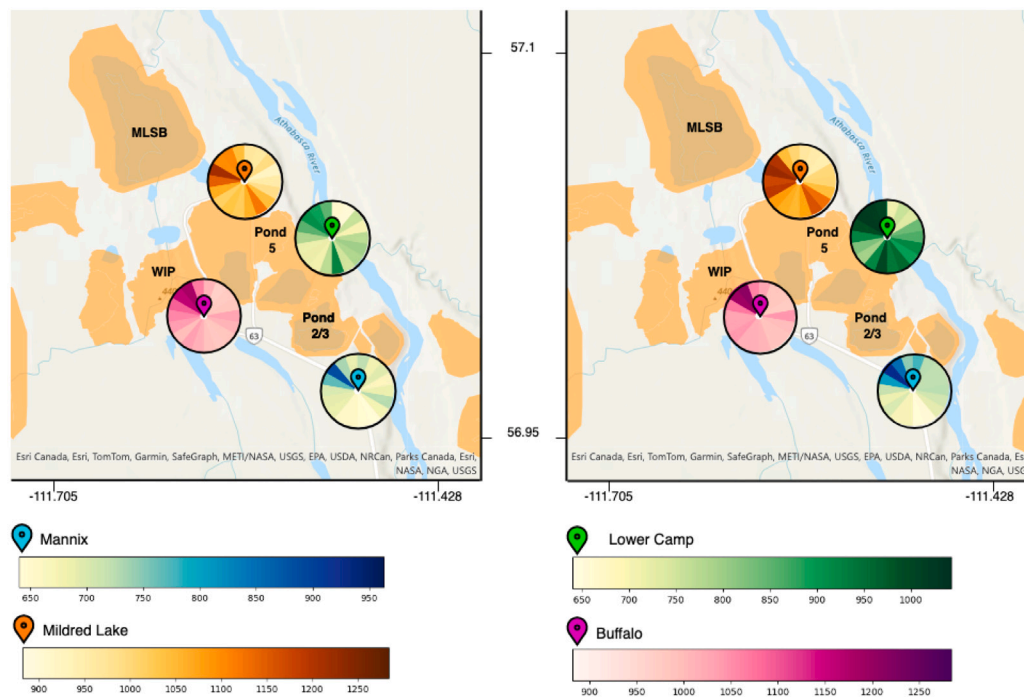


**Fig. 4.** Scaled values (between 0 and 1) of true versus predicted  $\text{CH}_4$  concentrations and emissions in 2023 for stations Mannix, Lower Camp and Mildred Lake using a neural network representation of the dispersion/advection terms in the ADM. The trained model accurately forecasts concentrations and emissions for one year when trained on historical data, suggesting future trends based on different levels of input data.

significant levels of emissions from Pond 5 can also be cross-verified with the results obtained for stations Lower Camp and Mildred Lake. While perfect sectoral disaggregation of sources is challenging, the choice of weather stations (close to OSTPs with no other  $\text{CH}_4$  source in between), use of radial wind directions as established in prior works and the fact that  $\text{CH}_4$  is generally measured within 5–15 m from the ground (indicating that  $\text{CH}_4$  does not disperse very far from its source,

thus the concentration recorded at the station has a source nearby) gives a strong indication of OSTPs being  $\text{CH}_4$  emission sources.

Upon comparing the plots for the years 2020 (left) and 2023 (right) in Fig. 5, we see that there has been a significant increase in emissions over the three years by at least 100 t in each of the 20 degree wind direction interval near the tailing ponds. Both MLSB and Pond 2/3 have started contributing higher amounts of  $\text{CH}_4$ . For year 2023, we also



**Fig. 5.** Estimation of yearly  $\text{CH}_4$  emissions (in t) from all wind directions around each station for 2020 (left) and 2023 (right). Estimated emissions are highest from the direction of OSTPs, with MLSB emitting 4500t per year and Pond 2/3 emitting more than 850 tonnes (t) per year every 20 degree interval. Abandoned lakes such as WIP and Pond 5 also emits more than 2000t of  $\text{CH}_4$  per year. Please note that the round plots are for demonstration purpose and are not to scale. Latitude and longitude coordinate ranges given in  $X$  and  $Y$  axis.

find elevated emissions for both Mannix and Lower Camp with wind directions 20–140 degrees and 60–140 degrees respectively, coming from OSTPs (such as Pond 8 owned by Suncor or similar sources in the area) south-east of the stations across the Athabasca river. Some other directions with lower emission levels of about 700–900 t (West of Mannix, South-East of Buffalo, SW of Lower Camp), all correspond to either in-situ facilities, industrial activity or inactive tailing ponds. Note that, the confidence interval for the emission predictions were similar to that obtained earlier in Fig. 3. However, since visualization of the interval is challenging in a radial plot, we have not included them in Fig. 5.

## 5. Discussion

Atmospheric  $\text{CH}_4$  can be affected by natural as well as anthropogenic factors. In order to build a reliable framework, DIRNN was trained by connecting all possible dynamics connected to  $\text{CH}_4$  emissions from OSTPs and the air quality in the region: MM of hydrocarbon degradation (Venegas Garcia, 2024; Kong et al., 2019), ADMs for  $\text{CH}_4$  dispersion (Stockie, 2011) and atmospheric data that drives atmospheric diffusion and advection. The two OSTPs used for dataset building are appropriate choices for representing majority of the OSTPs in the region mainly because of two reasons: (i) both of these ponds are not only big in terms of area but have also been previously found to emit much larger amounts of  $\text{CH}_4$  in comparison to other ponds; (ii) an important aspect of  $\text{CH}_4$  emissions from OSTPs is based on diluent composition which is determined by the owner company. Both these ponds take into account two of the major companies: Syncrude (now Suncor) and Suncor for our studies. These two ponds have been used in prior works frequently to study the extent of OSTP emissions (You et al., 2020; Small et al., 2015).

The model performances are compared based on both, plots and relative errors. Table 1 shows that for predicting concentrations, different parameterization of the  $\nabla \cdot J$  term can have similar predictive capabilities (Andoni et al., 2014). The differences in the model performances

between stations are attributed to the quality and quantity of data available for building the dataset. External factors such as the proximity of OSTPs to the weather stations, height of measurement sensors, etc. can affect data quality. For example, hypothetically if one of the stations were to be on a cliff (e.g. WBEA Stony Mountain, not used in the paper) it is more likely to be affected by high wind speeds (Solano et al., 2021), thus leading to a noisy dataset. Since the concentration data is noisy, we also see noise/oscillations in fitting of the emission data (also known as Gibbs phenomenon oscillations in literature) which is common in literature with model learning using noisy data (Berry and Harlim, 2016). Underlying model can be successfully recovered even with noisy data as long as the variation of noise is low (Tran and Ward, 2017) and the fitted function lies within the confidence of interval, thus controlling range within which the true values are expected to lie. Previous works have showed that Gibbs phenomenon does not majorly affect global generalizability of model approximation, and that different techniques of filtering illogical approximations (for example, negative values in a positive function approximation) is acceptable (Gottlieb and Shu, 1997; Berry and Harlim, 2016). The  $\text{CH}_4$  concentration data collected for station Mannix has the highest variance of 0.12, followed by station Mildred Lake with a variance of 0.05, and then station Lower Camp with a variance at 0.03. From the table of relative errors (Table 1), it can be seen that a lower the variance improves the performance of the models. This happens as data-driven models tend to have larger biases in predicting the extreme values, especially in data-scarce training regimes such as in this scenario. Superior model performance for the station Lower Camp can also be attributed to its restricted range of wind direction during the dataset building stage, making the input dataset more less noisy in comparison to other stations. Note that since the ponds are owned by different companies, the hydrocarbon degradation dynamics can differ based on the chemical composition of diluents used, as well as the accuracy of the data reported by the companies. Although emission predictions are slightly noisy, it does not affect the generalization power of our trained



**Table 2**

Comparison of CH<sub>4</sub> emission estimates for Pond 2/3 obtained from prior works and inverse dispersion modeling software WindTrax. Note that all the estimates are made for different years and may not be comparable directly. Units from original works are converted to tonnes per year for the 2.8 km<sup>2</sup> OSTP Pond 2/3.

| Name/Method             | Year      | Emission (t/y) |
|-------------------------|-----------|----------------|
| Small et al. (2015)     | 2015      | 2657.2         |
| You et al. (2020)       | 2016      | 3876.6         |
| You et al. (2020)       | 2018      | 11 344.2       |
| Windtrax (Crenna, 2016) | 2021      | 5518           |
| DIRNN                   | 2020–2023 | 3498.75        |

model as it preserves the overall trends on unseen data for both the quantities.

In the reverse formulation of the model, we use the trained model to monitor emissions from all the direction around each weather monitoring station. We consider the real concentration data along with the weather parameters inside the trained model to track emissions and possible sources (replace the input  $u_{\theta}$  in Eq. (6) with true CH<sub>4</sub> concentration data). Thus, for sources such as OSTPs (or other sources emitting CH<sub>4</sub> from diluent degradation) the model can be used as a tool to monitor emissions around weather stations. Since abandoned ponds and EPLs are often ignored as sources, our goal is also to monitor these tailings for emissions. A thorough analysis of emissions over four years in Figure C.27 shows that emissions from active OSTPs are not only increasing every year, some of the other inactive ponds/EPLs such as WIP and Pond 5 could also be significant sources of CH<sub>4</sub> emissions. Overall, our estimates show an underestimation in official reports (Canadian Centre for Energy Information, 2025; Government of Alberta, 2022) atleast by a factor of three.

### 5.1. Comparison to other works and data

In order to put our emission estimates into the context of existing literature, we take a look at the trend of emission estimates from previous works in Table 2. Note that since these studies were conducted across a span of last ten years, the emission estimates vary. In fact, they may also vary for different methods within the same year depending on the estimation method used. We include estimates using WindTrax (Crenna, 2016), a Lagrangian stochastic particle based model and calculated with parameters in You et al. (2020). Note that using WindTrax with limited weather station input data for 2020–2023 (such as from Lower Camp) leads to highly unstable results, with unrealistic outputs varying over 10<sup>5</sup>t per year. Thus, the outputs from those years are excluded from Table 2. Our estimates are based on emissions averaged over the four years 2020–2023 calculated from the station Lower Camp (or Mannix; both give similar estimates). CH<sub>4</sub> emissions using iDIRNN fall within the overall range of what is suggested by previous works. Our estimations of 3498.75 t/y are close to results showed by most of the estimations in You et al. (2020), Small et al. (2015). Note that due to different time periods, it may be hard to pinpoint which technique is the most accurate one.

The results underscore the need to increase efforts in accurately estimating CH<sub>4</sub> emissions and concentrations owing to its environmental and health impacts. Oil sands activities contribute significantly to GHG emissions and in particular are regarded as sources of pollution (Schindler, 2014; Liggio et al., 2016; Yu and Zahidi, 2023). From the perspective of environmental effects, mining and extraction of oil sands is directly associated with deforestation and release of sulfur oxides, nitrogen oxides, hydrocarbons, and fine particulate matter, etc. Further, CH<sub>4</sub> reacts with hydroxyl radical leading to the formation of ground-level ozone, which is a harmful and toxic air pollutant. While current levels of CH<sub>4</sub> concentration do not have direct health impact (~ 2 ppm–4.5 ppm), consequent displacement of oxygen, and ozone

formation can cause symptoms such rapid heart rate, fatigue and other health affects from lack of oxygen and increased air pollution (Adgate et al., 2014; Government of Canada, C. C. for O. H. and S., 2024).

Our study also aims to extend the growing body of research emphasizing on the role of machine learning in promoting sustainable practices, and efficient resource utilization. By integrating physics-informed deep learning into the monitoring and estimation of methane emissions from oil sands tailings ponds, our work contributes to the advancement of cost-effective environmental monitoring systems, offering tools for improving accountability, resource management, and regulatory compliance within the oil and gas sector industries. By making emission estimation scalable, accurate, and physically grounded, our methodology enhances decision-making capacity for governments, industries, and communities alike, promoting a fairer distribution of environmental responsibilities and better-informed climate governance.

### 5.2. Model uncertainties, limitations and future work

Being significantly data dependent, the training and results of our framework are limited to quality and quantity of available data. The data can be highly variable and prone to noise. For example, calibration errors, human recording errors, etc. The averaging step of processing the atmospheric data helps to ensure that noise in the data is reduced. The hybrid machine learning model also has a possibility of overfitting and unstable training. To avoid this, the results were based on the median values of multiple rounds of model training and validation with carefully tuned hyperparameters. Also note that the same trained model may not be applicable to other OSTPs without retraining. A model trained on one OSTP (and weather station) can be used on another location only if all the following conditions are met: (a) the OSTP under consideration is owned by the same company ensuring that the chemical composition of diluents remain the same; (b) it is located in a region with similar distribution of input parameters (similar weather conditions and/or diluent composition); (c) close proximity to the original OSTP and the weather monitoring station. For example, in order to get results for OSTPs owned by Shell or CNRL, we would need to retrain the model as (a) diluent composition of these companies are different; (b) they are far from the OSTPs we trained our model on. Since the model training is quick, if data is available, we would always recommend retraining to get the most accurate and reliable results. Future potential of this work is vast and includes improving MMs through advanced lab experiments, incorporating data from all directions around the weather stations to account for all possible CH<sub>4</sub> emitting sources and modeling the affects of these emissions on climate change.

## 6. Conclusions

In this paper, we developed a hybrid machine learning approach for predicting methane emissions and concentrations jointly. Our model formulation was based on learning the atmospheric methane concentrations using data obtained from weather monitoring stations subject to atmospheric dispersion models. We use a recurrent neural network style formulation to learn both, methane concentrations and the unknown functions in the constraints. Our input dataset included measurements from weather monitoring stations located within 4 km of active OSTPs and emission data obtained from solving different methanogenesis models. Along with providing accurate forecasts for CH<sub>4</sub> emissions and concentrations, our model results indicated significant emissions from both active and abandoned OSTPs, suggesting atleast three times underestimation in existing official reports. Our results using the proposed approach were consistent with existing estimates and outperformed other classical machine learning approaches. As a part of future work, we plan to incorporate sources from all the wind directions into the framework by either using additional constraints based on existing mathematical models of various methane sources or by incorporating aspects of remote sensing to gather data of methane emissions around the OSTPs.

## CRedit authorship contribution statement

**Esha Saha:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Oscar Wang:** Validation, Methodology, Data curation. **Amit K. Chakraborty:** Validation, Investigation, Data curation. **Pablo Venegas Garcia:** Writing – review & editing, Validation, Formal analysis, Data curation. **Russell Milne:** Writing – review & editing, Funding acquisition. **Hao Wang:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly and ChatGPT in order to improve writing in certain sections. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Esha Saha reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors would like to acknowledge the and thank the Natural Sciences and Engineering Research Council of Canada, Canada (NSERC) for funding through an NSERC Alliance Missions grant on anthropogenic greenhouse gas research.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jenvman.2025.127748>.

## Data availability

The data and code can be accessed at <https://github.com/eshasaha/champ>. This link has also been shared in the manuscript.

## References

- Adgate, J.L., Goldstein, B.D., McKenzie, L.M., 2014. Potential public health hazards, exposures and health effects from unconventional natural gas development. *Environ. Sci. Technol.* 48, 8307–8320.
- Alberta Energy Regulator, 2023. Alberta mineable oil sands plant statistics. <https://www.aer.ca/providing-information/data-and-reports/statistical-reports/st39>, Accessed 10 May 2024.
- Andoni, A., Panigrahy, R., Valiant, G., Zhang, L., 2014. Learning polynomials with neural networks. In: *International Conference on Machine Learning*. PMLR, pp. 1908–1916.
- Antonion, K., Wang, X., Raissi, M., Joshie, L., 2024. Machine learning through physics-informed neural networks: Progress and challenges. *Acad. J. Sci. Technol.* 9, 46–49.
- Bakels, L., Tatsii, D., Tipka, A., Thompson, R., Dütsch, M., Blaschek, M., Seibert, P., Baier, K., Bucci, S., Cassiani, M., et al., 2024. Flexpart version 11: Improved accuracy, efficiency, and flexibility. *Geosci. Model. Dev.* 17, 7595–7627.
- Basher, S.A., Sadorsky, P., 2025. How important are climate change risks for predicting clean energy stock prices? evidence from machine learning predictive modeling and interpretation. *J. Clim. Financ.* 10, 100058.
- Berry, T., Harlim, J., 2016. Semiparametric modeling: correcting low-dimensional model error in parametric models. *J. Comput. Phys.* 308, 305–321.
- Bonifacio, H.F., Maghirang, R.G., Razote, E.B., Trabue, S.L., Prueger, J.H., 2013. Comparison of aermid and windtrax dispersion models in determining pm10 emission rates from a beef cattle feedlot. *J. Air Waste Manage. Assoc.* 63, 545–556.
- Burkus, Z., Wheler, J., Pletcher, S., 2014. Ghg emissions from oil sands tailings ponds: Overview and modelling based on fermentable substrates part i: Review of the tailings ponds facts and practices part ii: Modelling of ghg emissions from tailings ponds based on fermentable substrates zvonko burkus alberta environment and sustainable resource development (aesrd).
- Canadian Centre for Energy Information, 2025. Greenhouse gas emissions — national inventory report (nir). <https://energy-information.canada.ca/en/subjects/greenhouse-gas-emissions-national-inventory-report-nir>. (Accessed 15 January 2025).
- Cheng, H., Hamidoğlu, A., Sysoeva, L., Garcia, P.V., Milne, R., Burkus, Z., Wang, H., 2025. A novel evolutionary game-based low-methane application in three-echelon energy supply chains. *Appl. Energy* 401, 126777.
- Cimorelli, A.J., Perry, S.G., Venkatram, A., Weil, J.C., Paine, R.J., Wilson, R.B., Lee, R.F., Peters, W.D., Brode, R.W., 2005. Aermid: A dispersion model for industrial source applications. part i: General model formulation and boundary layer characterization. *J. Appl. Meteorol.* 44, 682–693.
- Crenna, B., 2016. An introduction to windtrax. *J. Env. Prot* 7.
- Das, D.B., Tatarwal, S.K., Sharma, R., 1998. Release using dispersion modeling. *Q. J. Hung. Meteorol. Serv.* 102, 167–187.
- D'Orazio, P., Pham, A.D., 2025. Evaluating climate-related financial policies' impact on decarbonization with machine learning methods. *Sci. Rep.* 15 (1694).
- Flannigan, M.D., Krawchuk, M.A., de Groot, W.J., Wotton, B.M., Gowman, L.M., 2009. Implication of changing climate for global wildland fire. *Int. J. Wildland Fire* 18 (5), 483–507. <http://dx.doi.org/10.1071/WF08187>.
- Foght, J.M., Gieg, L.M., Siddique, T., 2017. The microbiology of oil sands tailings: past, present, future. *FEMS Microbiol. Ecol.* 93, fix034.
- Gosselin, P., Hrudey, S.E., Naeth, M.A., Plourde, A., Therrien, R., Van Der Kraak, G., Xu, Z., 2010. Environmental and Health Impacts of Canada's Oil Sands Industry. Royal Society of Canada, Ottawa, ON, p. 10.
- Gottlieb, D., Shu, C.W., 1997. On the gibbs phenomenon and its resolution. *SIAM Rev.* 39, 644–668.
- Government of Alberta, 2022. Area fugitive emissions from oil sands mines. <https://open.alberta.ca/opendata/area-fugitive-emissions-from-oil-sands-mines>. Accessed 10 September 2024.
- Government of Canada, C. C. for O. H. and S., 2024. Methane. [https://www.ccohs.ca/oshanswers/chemicals/chem\\_profiles/methane.html](https://www.ccohs.ca/oshanswers/chemicals/chem_profiles/methane.html). Accessed 10 April 2025.
- Hamrani, A., Akbarzadeh, A., Madramootoo, C.A., 2020. Machine learning for predicting greenhouse gas emissions from agricultural soils. *Sci. Total Environ.* 741, 140338.
- Hou, L., Dai, Q., Song, C., Liu, B., Guo, F., Dai, T., Li, L., Liu, B., Bi, X., Zhang, Y., et al., 2022. Revealing drivers of haze pollution by explainable machine learning. *Environ. Sci. Technol. Lett.* 9, 112–119.
- Hu, L., Wang, C., Ye, Z., Wang, S., 2021. Estimating gaseous pollutants from bus emissions: A hybrid model based on gru and xgboost. *Sci. Total Environ.* 783, 146870.
- Islam, M.R., Islam, M.M., Badhan, I.A., Hasnain, M.N., 2025. The role of artificial intelligence in carbon pricing policies: Economic and environmental implications. *J. Eng. Comput. Intell. Rev.* 3.
- Jin, B., Xu, X., 2024. Price forecasting through neural networks for crude oil, heating oil, and natural gas. *Meas.: Energy* 1, 100001.
- Jin, B., Xu, X., 2024a. Forecasts of thermal coal prices through gaussian process regressions. *Ironmak. Steelmak.* 51, 819–834.
- Jin, B., Xu, X., 2025. Bayesian gaussian process predictions of chongqing carbon market prices. *J. Environ. Assess. Policy Manag.* 2550013.
- Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmailzadeh, S., Azizzadehsheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al., 2021. Physics-informed machine learning: case studies for weather and climate modelling. *Phil. Trans. R. Soc. A* 379, 20200093.
- Kong, J.D., Wang, H., Siddique, T., Foght, J., Semple, K., Burkus, Z., Lewis, M.A., 2019. Second-generation stoichiometric mathematical model to predict methane emissions from oil sands tailings. *Sci. Total Environ.* 694, 133645.
- Koushafar, M., Sohn, G., Gordon, M., 2023. Deep convolutional neural network for plume rise measurements in industrial environments. *Remote. Sens.* 15 (3083).
- Kumar, S., Patel, V., Mehta, J., Verma, J., Sharma, A.K., 2025. A comparison of statistical and machine learning models for stock price prediction. *Int. J. Inf. Decis. Sci.* 17, 168–193.
- Liggio, J., Li, S.M., Hayden, K., Taha, Y.M., Stroud, C., Darlington, A., Drollette, B.D., Gordon, M., Lee, P., Liu, P., et al., 2016. Oil sands operations as a large source of secondary organic aerosols. *Nature* 534, 91–94.
- Luo, R., Wang, J., Gates, I., 2023. Machine learning for accurate methane concentration predictions: short-term training, long-term results. *Environ. Res. Commun.* 5, 081003.
- McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133.
- Meng, X., Chang, H., Wang, X., 2022. Methane concentration prediction method based on deep learning and classical time series analysis. *Energies* 15 (2262).

- Michel, S.E., Lan, X., Miller, J., Tans, P., Clark, J.R., Schaefer, H., Sperlich, P., Brailsford, G., Morimoto, S., Moossen, H., et al., 2024. Rapid shift in methane carbon isotopes suggests microbial emissions drove record high atmospheric methane growth in 2020–2022. *Proc. Natl. Acad. Sci.* 121, e2411212121.
- Mikkelsen, T., Larsen, S., Pécseli, H., 1987. Diffusion of gaussian puffs. *Q. J. R. Meteorol. Soc.* 113, 81–105.
- Mohan, M., Siddiqui, T., 2002. Development of an atmospheric dispersion model for air quality assessment. In: 8th Int. Conf. on Harmonisation Within Atmospheric Dispersion Modelling for Regulatory Purposes.
- Penner, T.J., Foght, J.M., 2010. Mature fine tailings from oil sands processing harbour diverse methanogenic communities. *Can. J. Microbiol.* 56, 459–470.
- Psomouli, T., Kansizoglou, I., Gasteratos, A., 2023. Methane concentration forecasting based on sentinel-5p products and recurrent neural networks. *Geosciences* 13 (183).
- Qu, Z., Jacob, D.J., Bloom, A.A., Worden, J.R., Parker, R.J., Boesch, H., 2024. Inverse modeling of 2010–2022 satellite observations shows that inundation of the wet tropics drove the 2020–2022 methane surge. *Proc. Natl. Acad. Sci.* 121, e2402730121.
- Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378, 686–707.
- Ramos-Padrón, E., Bordenave, S., Lin, S., Bhaskar, I.M., Dong, X., Sensen, C.W., Fournier, J., Voordouw, G., Gieg, L.M., 2011. Carbon and sulfur cycling by microbial communities in a gypsum-treated oil sands tailings pond. *Environ. Sci. Technol.* 45, 439–446.
- Saha, E., Ho, L.S.T., Tran, G., 2023a. Spade4: Sparsity and delay embedding based forecasting of epidemics. *Bull. Math. Biol.* 85 (71).
- Saha, E., Schaeffer, H., Tran, G., 2023b. Harfe: hard-ridge random feature expansion. *Sampl. Theory, Signal Process. Data Anal.* 21 (27).
- Saha, Esha, Tran, Giang, 2023. Generalization bound for diffusion models using random features. *arXiv preprint arXiv:2310.04417*.
- Schaeffer, H., Tran, G., Ward, R., 2018. Extracting sparse high-dimensional dynamics from limited data. *SIAM J. Appl. Math.* 78, 3279–3295.
- Schindler, D.W., 2014. Unravelling the complexity of pollution by the oil sands industry. *Proc. Natl. Acad. Sci.* 111, 3209–3210.
- Siddique, T., Fedorak, P.M., Foght, J.M., 2006. Biodegradation of short-chain n-alkanes in oil sands tailings under methanogenic conditions. *Environ. Sci. Technol.* 40, 5459–5464.
- Siddique, T., Fedorak, P.M., MacKinnon, M.D., Foght, J.M., 2007. Metabolism of btx and naphtha compounds to methane in oil sands tailings. *Environ. Sci. Technol.* 41, 2350–2356.
- Siddique, T., Gupta, R., Fedorak, P.M., MacKinnon, M.D., Foght, J.M., 2008. A first approximation kinetic model to predict methane generation from an oil sands tailings settling basin. *Chemosphere* 72, 1573–1580.
- Small, C.C., Cho, S., Hashisho, Z., Ulrich, A.C., 2015. Emissions from oil sands tailings ponds: Review of tailings pond parameters and emission estimates. *J. Pet. Sci. Eng.* 127, 490–501.
- Solano, J., Montañó, T., Maldonado-Correa, J., Ordóñez, A., Pesantez, M., 2021. Correlation between the wind speed and the elevation to evaluate the wind potential in the southern region of Ecuador. *Energy Rep.* 7, 259–268.
- Stockie, J.M., 2011. The mathematics of atmospheric dispersion modeling. *Siam Rev.* 53, 349–372.
- Sysoeva, L., Bouderbala, I., Kent, M.H., Saha, E., Zambrano-Luna, B., Milne, R., Wang, H., 2025. Decoding methane concentration in Alberta oil sands: A machine learning exploration. *Ecol. Indic.* 170, 112835.
- Tagliaferri, F., Invernizzi, M., Sironi, S., et al., 2022. Validation study of windtrax backward lagrangian model. In: Harmo 21 Proceedings. Universidade de Aveiro, pp. 463–468.
- Tong, W., Li, L., Zhou, X., Hamilton, A., Zhang, K., 2019. Deep learning pm 2.5 concentrations with bidirectional lstm rnn. *Air Qual. Atmosphere & Heal.* 12, 411–423.
- Tran, G., Ward, R., 2017. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Model. Simul.* 15, 1108–1129.
- US EPA, 2024. Overview of greenhouse gases. <https://www.epa.gov/ghgemissions/overview-greenhouse-gases#methane>. Accessed 09 May 2024.
- Venegas Garcia, P., 2024. Modelling microbial processes of soil organic matter mineralization and methane production. PhD Dissertation Thesis.
- Vojta, M., Plach, A., Thompson, R.L., Stohl, A., 2022. A comprehensive evaluation of the use of lagrangian particle dispersion models for inverse modeling of greenhouse gas emissions. *Geosci. Model. Dev.* 15, 8295–8323.
- Wood Buffalo Environmental Association, 2024. Network map & station data. <https://wbea.org/data/network-map-station-data/>. (Accessed 15 May 2024).
- Xie, H., Ji, L., Wang, Q., Jia, Z., 2019. Research of pm2.5 prediction system based on cnns-gru in wuxi urban area. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing, 032073.
- Xu, K., Darve, E., 2022. Physics constrained learning for data-driven inverse modeling from sparse observations. *J. Comput. Phys.* 453, 110938.
- Yang, J., Sun, Y., Li, Q., Sun, Y., 2020. Effective risk prediction of tailings ponds using machine learning. In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering. AEMCSE, IEEE, pp. 234–238.
- Yazdinejad, A., Wang, H., Kong, J., 2025. Advanced ai-driven methane emission detection, quantification, and localization in Canada: A hybrid multi-source fusion framework. *Sci. Total Environ.* 998, 180142.
- You, Y., Staebler, R.M., Moussa, S.G., Beck, J., Mittermeier, R.L., 2020. Methane emissions from an oil sands tailings pond: a quantitative comparison of fluxes derived by different methods. *Atmospheric Meas. Tech. Discuss.* 2020, 1–24.
- Yu, H., Zahidi, I., 2023. Tailings pond classification based on satellite images and machine learning: An exploration of microsoft ml. net.. *Mathematics* 11 (517).
- Zambrano-Luna, B.A., Sysoeva, Liubov, Gao, Shan, Milne, Russell, Burkus, Zvonko, Wang, Hao, 2025. Improved monitoring of methane emissions for the oil and gas sector with Sentinel-2 satellite observations. *Atmos. Environ.* 121594.