Reducing Air Pollution through Machine Learning

Dimitris Bertsimas[a,b], Léonard Boussioux[a,b], Cynthia Zeng[a,b]

[a] *Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA*
[b] *Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA*

Abstract

Problem Definition: This paper presents a data-driven approach to mitigate the effects of air pollution from industrial plants on nearby cities by linking operational decisions with weather conditions.

Academic Relevance: Our method combines predictive and prescriptive machine learning models to forecast short-term wind speed and direction and recommend operational decisions to reduce or pause the industrial plant's production. We exhibit several trade-offs between reducing environmental impact and maintaining production activities.

Method and Results: The predictive component of our framework employs various machine learning models, such as gradient-boosted tree-based models and ensemble methods, for time series forecasting. The prescriptive component utilizes interpretable optimal policy trees to propose multiple trade-offs, such as reducing dangerous emissions by 33-47% and unnecessary costs by 40-63%. Our deployed models significantly reduced forecasting errors, with a range of 38-52% for less than 12-hour lead time and 14-46% for 12 to 48-hour lead time compared to official weather forecasts. We have successfully implemented the predictive component at the OCP Safi site, which is Morocco's largest chemical industrial plant, and are currently in the process of deploying the prescriptive component.

Managerial Insights: Our framework provides a pathway for sustainable industrial development by forgoing the trade-off between pollution and industrial activity by linking operational decisions with data-driven weather conditions. This represents a significant step in optimizing factory operations and improving sustainability efforts. As such, it has the

potential to modernize how factories approach planning and resource allocation under environmental compliance. Our predictive component has significantly improved production efficiency, allowing for better resource allocation and reduced downtime. This not only led to cost savings for the company but also helped to reduce the environmental impact of production by minimizing air pollution.
Keywords: Air Pollution Management, Machine Learning, Sustainability, Predictive and Prescriptive Analytics, Plant Operations

1. Introduction

Sustainable industrial development is an important issue shared by many countries. The trade-offs between economic activities, environmental pollution, and public health must be managed attentively. Studies show that urbanization and industrialization have released many environmental toxins into the atmosphere over the last 200 years [1, 2, 3]. In particular, emissions from chemical power plants can pose significant health risks to those living in the surrounding area [4, 5]. Therefore, there is a pressing need to develop technologies and infrastructures to simultaneously achieve economic objectives and environmental preservation.

As data availability and computing methods continue to advance, there has been growing interest in applying machine learning techniques to air pollution management. Previous research has primarily focused on predicting the health consequences of pollution exposure [6]. Additionally, various studies have attempted to forecast air pollution, air quality, and airborne particle concentrations using data such as satellite imagery, weather data, and air quality monitoring data [7, 8, 9, 10, 11]. Despite these efforts, there remains a lack of literature connecting air pollution prediction to decision-making and mitigation actions. Earlier works on technology-aided tools to reduce pollution include [12], which discusses a mathematical formulation and algorithm for controlling air pollution using weather forecasts

and numerical models to minimize control-related costs, and [13], which proposes a decision support tool to find optimal Best Management Practice locations for minimizing diffuse surface water pollution.

This paper tackles the critical issue of urban air pollution management by proposing a novel plant operation scheduling methodology that leverages machine learning and optimization. Our predictive and prescriptive framework links operational decisions to weather forecasts to effectively minimize the impact of air pollution in industrial settings. To the best of our knowledge, our work is the first attempt to reduce industrial air pollution through machine learning. In addition, the framework is implemented and currently operational on the Safi production site of the OCP group in Morocco. In summary, our contributions are three-fold:

- A data-driven pollution framework incorporating two components: (i) a machine learning-enhanced weather forecasting system that utilizes onsite sensors and official forecasts (ii) an optimization-based operational decision recommendation system optimizing the trade-off between potential pollution risk and operational loss. The predictive component of our framework has been deployed and guides production planning in real-time at the OCP Safi site since July 2021. The prescriptive component is under implementation.

- Since implementation, our machine learning-enhanced forecasts significantly improved accuracy: we reduced the next 12-hour wind forecasting errors by 38-52% and the next 12 to 48-hour errors by 14-46%. In addition, our optimization-based operational decision framework is shown to reduce potential polluting cases by 33-47% while achieving 40-63% operational savings.

- Our work offers a case study of achieving industrial activities while controlling air pollution's impact on surrounding urban cities. We hope to inspire future work applying

machine learning and data science for sustainable industrial development.

2. Methodology

2.1. The Previous Operational Procedure at Safi

The OCP group is the world's largest phosphate producer, controlling 75% of the world's phosphate reserves and accounting for more than 30% of global production. The OCP Safi site was established in the 1970s to produce various phosphates for export. However, fertilizer production is a known contributor to air pollution, releasing harmful airborne substances such as sulfur dioxide (SO_2), sulfur trioxide (SO_3), hydrogen sulfide (H_2S), and hydrogen fluoride (HF), as well as fine and coarse dust, which can pose serious health risks such as respiratory diseases and cancer [14]. The site is located 10 km southwest of the Safi city center, with more than 300,000 residents. Due to the geographical location, weather conditions play a critical role in air pollution dispersion. Depending on the wind speed and direction, airborne pollution can be carried into Safi, thus posing a threat to public health and bringing high respiratory and ocular discomfort. In 2013, the site set up a monitoring procedure to reduce the amount of air pollution in the city with responsive production rates — and consequently airborne emissions — depending on the meteorological weather forecasts and real-time on-site wind monitoring system. This procedure schedules production rates and personnel based on next-day weather forecasts. It uses real-time wind monitoring systems to adjust in dangerous weather conditions, ensuring the safety of the surrounding community.

Before this study, the main bottleneck of the procedure was the gap between meteorological forecasts and real-time conditions, thus leading to unnecessary and costly production shutdowns or missed dangerous weather conditions, leading to negative health outcomes. The operators in the Safi production site received operational weather forecasts from the national meteorological agency every 12 hours for the next 48 hours. However, these fore-

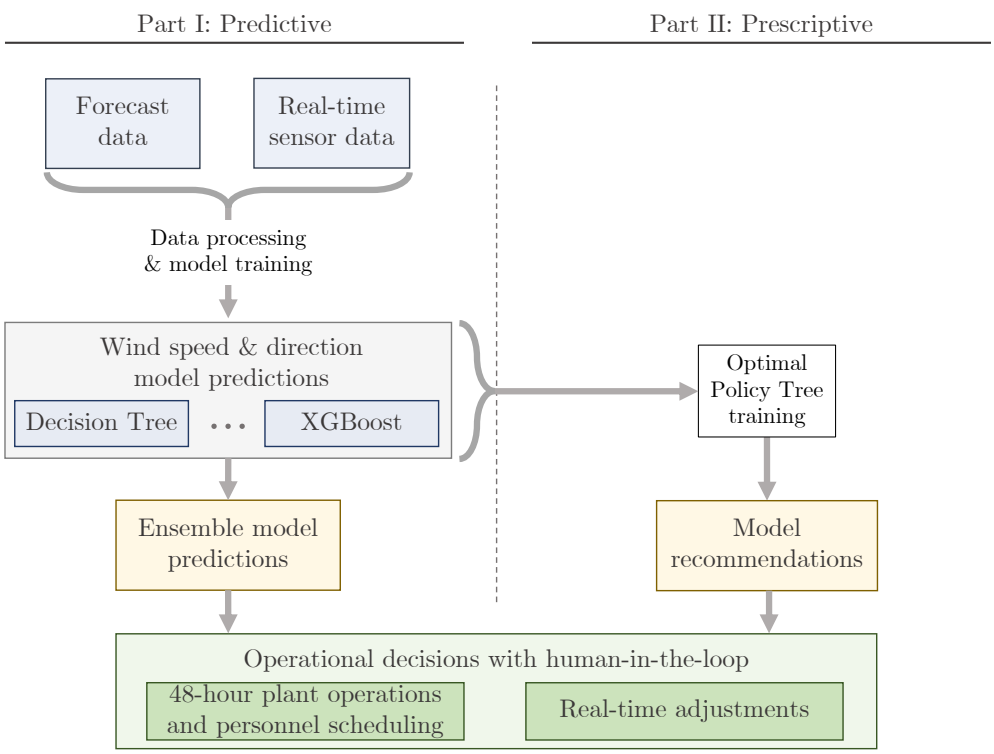

Figure 1: Predictive and prescriptive approach to plant operations scheduling.

casts are frequently inaccurate because they are calculated at a regional level and come with a 5 to 7-hour lag-time due to the long computational costs of dynamical weather forecasts. As a result, planning activities had no access to real-time forecast information and were sensitive to uncertain weather conditions.

This study aims to develop a data-driven framework to reduce the impact of air pollution from industrial plants on nearby cities by responsively adapting production levels based on wind speed and direction. Our pipeline encompasses two parts: i) machine learning algorithms producing more accurate and frequent wind forecasts by combining official weather data and onsite real-time sensory data to aid short-to-medium term factory and personnel planning; ii) an optimization-based framework to recommend real-time optimal operational decisions taking into account the various forecasts from the machine learning models. Figure 1 illustrates our overarching methodology.

2.2. Scenario Definition

The OCP Safi site developed a warning system to categorize weather conditions into several scenarios to monitor the potential risk of wind carrying pollutants into the nearby city of Safi. Scenarios are differentiated as either *favorable* (S1, S2, S2b, S3) or *dangerous* (S3b, S4) based on wind speed and direction, as outlined in Table 1. This categorization accounts for five wind speed buckets and three wind direction buckets, providing detailed guidance on production rates for each scenario. A dangerous scenario is characterized by low wind speed combined with an unfavorable wind direction, which results in pollutants being directed toward and lingering in the city (see illustration in Figure 2). Based on the real-time and predicted scenarios, operational decisions are made to reduce air pollution according to the action rules outlined in Table 2.

Wind Speed ($m.s^{-1}$)	**Favorable Wind Direction** NW, N-NW, N, N-NE, NE, E-NE, E $0° - 101.25°$ & $303.75° - 0°$	**Very Unfavorable Wind Direction** S-SW, S, S-SE $146.25° - 213.75°$	**Unfavorable Wind Direction** E-SE, SE, SW, W-SW, W, W-NW $101.25° - 146.25°$ & $213.75° - 303.75°$
$V < 0.5$	S3 a	S4	S4
$0.5 \leq V < 1$	S2	S3 b	S2 b
$1 < V \leq 2$	S1	S3 b	S2 b
$2 < V \leq 4$	S1	S3 b	S2
$4 < V$	S1	S1	S1

Table 1: Scenario definitions based on wind speed and direction, accounting for five wind speed buckets and three wind direction buckets. Scenarios are differentiated as either favorable (S1, S2, S2b, S3a) or dangerous (S3b, S4).

Scenario Type	**Underlying Scenarios**	**Scenario Characteristics**	**Public Health Consequences**
Favorable	S1, S2, S2b, S3	High wind speed and/or favorable wind direction	Limited
Dangerous	S3b, S4	Low wind speed and unfavorable wind direction	Pollutants directed toward and lingering in city

Table 2: Categorization of scenarios as favorable and dangerous based on wind speed and direction.

The effectiveness of the plant's operational response to dynamic weather conditions is contingent on the accuracy of real-time forecasts. If a dangerous scenario is projected in the next three hours, the plant operators proactively adjust production levels accordingly. Once at reduced capacity, production remains at this level until a favorable scenario is predicted and real-time weather conditions become favorable. This highlights the critical importance

Figure 2: Wind direction and wind speed determine the dissemination of pollutants. Winds coming from the South with low speeds are the most dangerous conditions.

of precise weather forecasting, as inaccuracies can result in costly production and personnel scheduling consequences.

Before our work, plant managers used official regional forecasts as mere guidance and often relied on experience due to the low accuracy in near-term predictions. This led to frequent inconsistencies and last-minute adjustments under unforeseen weather conditions, underscoring the need for real-time, high-quality weather forecasts in informed decision-making.

2.3. Predictive Methodology

Our predictive framework focuses on developing machine learning models to produce accurate hourly wind forecasts by integrating official weather data and onsite real-time sensory data to aid short-to-medium factory and personnel planning. Since July 2021, our predictive framework has been implemented at the OCP Safi site, resulting in reduced production downtime, improved resource allocation, and cost savings. This successful implementation

serves as a model for other factories seeking to improve their sustainability efforts and reduce their environmental impact.

Data Processing. We combined two datasets to make predictions: the official regional weather forecast data and real-time weather measurement data collected with on-site sensors. Data used in this study range from July 2015 to March 2022.

Official forecasts are received twice daily, around 6:00 am (GMT) and 6:00 pm (GMT) from the Moroccan National Meteorological Department. These forecasts are produced by traditional dynamical models with initial conditions and often take 5-7 hours of computational time. They provide hourly values for the next 48 hours for wind speed, wind direction, humidity, solar irradiance, and temperature at the Safi site. We call this model the *baseline model* in the rest of the paper. The on-site sensors measure the same five weather features (wind speed, wind direction, humidity, solar irradiance, and temperature) at one-minute intervals.

We first imputed the missing values caused by electronic or server malfunctions with linear interpolation. We then averaged the measurement data over one-hour intervals. We used the arithmetic average for the humidity, solar irradiance, and temperature, and the vector average technique [15] for wind speed and direction (e.g., the vector average of a southerly and a northerly wind of 5 m.s^{-1} gives a mean wind speed of 0 m.s^{-1} because there is no resultant wind speed). We encoded the wind direction using the cosine and sine transformations to avoid singularities at endpoints due to the cyclical nature of the feature.

Training data creation. We transformed the time-series data into a standard tabular form to train traditional machine-learning models. To make wind predictions at time t for the hour $t+h$, we concatenated the present and past 48 hours of weather measurement features at each time step into a vector. Then, we appended the following features: the latest operational forecast available at time t for wind speed, wind direction, pluviometry, and solar irradiance;

the cosine and sine of the day and the hour corresponding to time t. Table 3 summarizes the 304 features and associated processing techniques.

Feature Description	**Processing Technique**	**Initial Feature Range**	**Number of features**
Wind speed	Vector average	0.00 - 14.20 $m.s^{-1}$	49
Wind direction	Vector average, cos/sin encoding	[0, 360]°	49×2
Solar irradiance	Arithmetic average	0.0 - 978.4 $W.m^{-2}$	49
Temperature	Arithmetic average	4.8 - 46.7°C	49
Pluviometry	Arithmetic average	0.0 - 17.2 mm	49
Day of the year	Cos/sin encoding	1 - 365	2
Hour of the day	Cos/sin encoding	0 - 23	2
Official forecast for wind speed		0.0 - 16.5 $m.s^{-1}$	1
Official forecast for wind direction	Cos/sin encoding	[0, 360]°	2
Official forecast for pluviometry		0.0 - 20.8 mm	1
Official forecast for solar irradiance		0 - 1074 $W.m^{-2}$	1
Official forecast for temperature		3.2 - 43.1°C	1

Table 3: Table recording all the features and processing techniques. The number of features obtained accounts for concatenating the past 48-hour values.

Model Training. For the prediction task, we trained five different types of machine learning models to predict wind speed and direction, including Elastic Net, Decision Trees, Random Forest, LightGBM, and XGBoost. To handle the cyclical property for wind direction, we predicted the cosine and sine of the angle instead of the raw angle degree. Predictions are then converted back into scenario predictions using Table 1. We trained one model for each lead time between 1 and 48 hours ahead, i.e., 48×3 regression models for wind speed, cosine, and sine of wind direction. We performed hyperparameter tuning for each model using the validation set as explained later in Section 2.5.

In addition, we trained ensemble models to predict wind speed and direction for every lead time using predictions from these previous individual machine-learning models. Ensemble modeling is a well-established technique to leverage the strengths and limitations of multiple models and benefit from their diversity. The principle is to combine the predictions of the forecasting models available to obtain a more accurate, stable, and robust predictor. In our case, we used the stacking [16] concept and tried several ensemblers, including deci-

sion trees, regularized linear regression, and gradient-boosted trees. Elastic Net regression performed the best, and we considered it our final ensemble model technique.

2.3.1. Qualitative Feedback from Real-World Implementation

Our collaboration with OCP's software development team has seamlessly integrated our weather forecasts into the company's internal system (see Figure 3). As of July 2021, the site manager and plant operators have been utilizing the forecasts produced by our framework through a simple user interface. They check the hourly forecasts before scheduling production shutdowns, leading to a significant reduction in production downtime.

Qualitative feedback from production managers has indicated that our forecasts are substantially more accurate than official weather forecasts and provide valuable real-time updates that are particularly advantageous during winter when wind conditions are more unpredictable. This has improved factory planning and resource allocation, allowing for more efficient production, better personnel scheduling, and cost savings for the company.

The successful implementation of our framework at OCP Safi is a testament to our approach's effectiveness in optimizing factory operations. We believe that utilizing our framework has the potential to advance how factories approach planning and resource allocation, ultimately leading to improved sustainability efforts and environmental impact reduction.

2.4. Prescriptive Methodology

The management team at OCP Safi recognizes the importance of taking immediate action in response to dangerous weather conditions. As a result, our focus is on utilizing short-term (next 3 hours) weather predictions to inform plant operations. We employed Optimal Policy Trees (OPT) [17] to determine the most optimal decision in real-time given the forecasts made by the different machine learning models. Despite the imbalanced nature of the data, with dangerous scenarios accounting for only 1.5-2% of the total observations, our prescriptive models aim to balance the trade-off between financial savings and effective

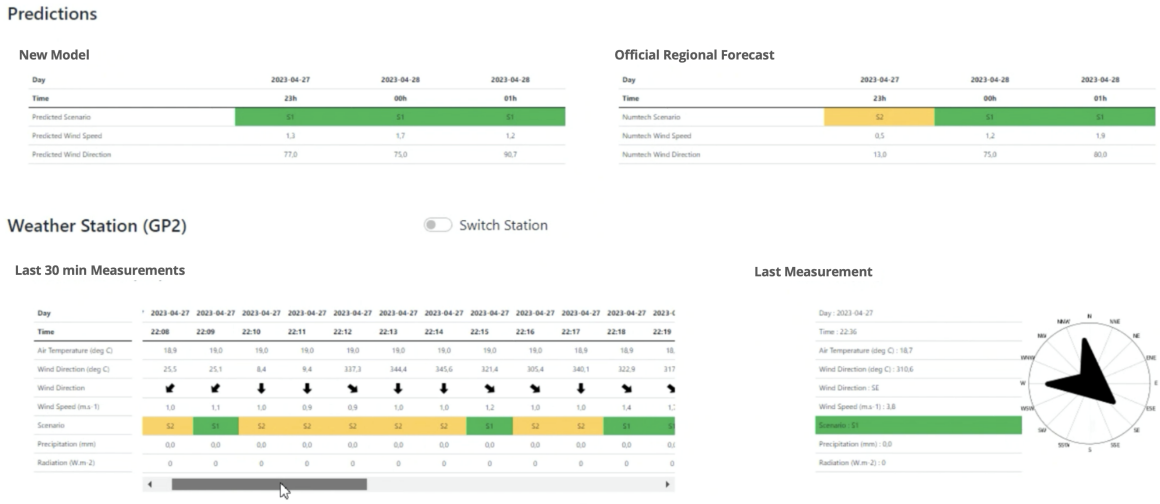

Figure 3: Screenshot of the software platform used by the plant operators in Safi. Our model predictions for the next 3 hours are displayed on the upper left, while the baseline model is displayed on the upper right. Below, the operator can check the real-time 1-min measurements of the previous 30 minutes to adapt decisions.

pollution management. This approach is crucial to mitigate the potential for false negatives in predictive models.

In our context, the prescriptive approach deals with observational data of the form $\{(\mathbf{x}_i, y_i, z_i)\}$. Each observation i consists of features $\mathbf{x}_i \in \mathbb{R}^{18}$ (the ensemble members' predictions), an applied prescription $z_i \in \{0, 1\}$ (reduce plant production or not), and an observed outcome $y_i \in \mathbb{R}$ (real-world costs associated with the decision). Our prescriptive task is determining the optimal policy that, given the features $\mathbf{x}$, prescribes the treatment z that results in the best outcome y. The prescription involves choosing between one of two available decisions, either to reduce production or not.

Table 4 outlines the reward matrix used to train the Optimal Policy Tree and quantifies the costs associated with false positives and false negatives. First, no cost is incurred if the forecasted scenario and actual conditions are favorable. When the plant operates at reduced levels as a conservative measure after forecasting a dangerous scenario, the factory incurs a loss of earnings of \$2,000 per hour due to decreased production and the expenses of injecting odor control chemicals to minimize unpleasant odors in the surrounding area. On the other

Forecasted Scenario	Actual Scenario	Cost (USD)	Decision Outcome	Public Health Impact
Favorable	Favorable	0	Full level production	Low
Dangerous	Favorable	2000	Reduced level production + anti-odor injection	Low
Dangerous	Dangerous	2000	Reduced level production + anti-odor injection	Low
Favorable	Dangerous	4000 - 20000	Full production before urgent shutdown + anti-odor injection	High

Table 4: Reward matrix for training the Optimal Policy Trees based upon forecasted and actual weather conditions.

hand, the failure to forecast a dangerous scenario leads to the plant operating at a normal level and polluting the nearby city when the weather conditions turn dangerous. Afterward, the plant operators must shut down production urgently and inject odor control chemicals. We propose evaluating various public health costs ranging from $2,000 to $18,000. This parameter yields differing trade-offs between pollution and costs and can be determined based on the decision-makers' conservatism and risk aversion level.

2.5. Training Protocol

The data covers August 2015 to March 2022, totaling 43,952 hourly samples. The data set was divided into training (60%), validation (20%), and testing (20%) sets. The validation set was used to tune the hyperparameters of the machine-learning models. The ensemble models and optimal policy tree parameters were 5-fold cross-validated on the predictions made on the validation set. All models were evaluated on the unseen test set corresponding to the real-world deployment phase.

Software Tools. We used Python 3.8 [18] and the scikit-learn package [19] to implement all machine learning models. We used the Python package InterpretableAI [20] to train Optimal Policy Trees.

3. Results

This section reports the results of the two components of the framework: predictive and prescriptive. We have successfully implemented the predictive component on machine

Model	Hyperparameters	Values
Elastic Net	regularization α coefficient	0.2, 0.4, 0.6, 0.8, 1
	ℓ_1 ratio	0.5, 0.7
Decision Trees	maximum tree depth	5, 6, 7, 8, 9, 10
	minimum samples per split	3, 5, 7
	minimum samples per leaf	4, 6
Random Forest	bootstrap	True, False
	number of estimators	100, 150
	maximum tree depth	5, 6
	min samples split	4, 6
LightGBM	number of leaves	31, 60
	maximum tree depth	4, 6
	learning rate	0.1, 0.3
	lambda ℓ_1	0, 1
XGBoost	number of estimators	100, 150
	maximum tree depth	4, 6
	learning rate	0.1, 0.3
Elastic Net Ensemble	regularization α coefficient	0.2, 0.4, 0.6, 0.8, 1
	ℓ_1 ratio	0, 0.25, 0.5, 0.75, 1.0

Table 5: Hyperparameters searched for our models.

learning-based wind forecasts since December 2020, and we are currently implementing the prescriptive component on operational decision-making recommendations. As such, we report real-world deployment results for the predictive component and back-tested results for the prescriptive component.

3.1. Predictive Methodology Results

Tables 6 and 7 report the results of the wind speed and wind direction forecasting tasks for all the regression models we deployed at the Safi site: the baseline model, Elastic Net, Decision Tree, Random Forest, Light GBM, XGBoost, and the Elastic Net ensemble model. For each wind prediction task, we report the mean absolute error (MAE) and the expected shortfall at 85%, corresponding to the average error on the worst 15% samples. The baseline model refers to the weather forecast guidance from the Moroccan Meteorological Department.

All machine learning models generally improve upon the baseline model, with XGBoost and Light GBM achieving the lowest errors. In addition, the ensemble model further improves the MAE and expected shortfall, especially in near-term horizons. Looking at speed prediction, for less than 12-hour lead time, the best-performing machine learning approaches can outperform the baseline model by 40-50% in both metrics. For longer-term predictions with more than a 12-hour horizon, the best-performing machine learning approaches can outperform the baseline model by 20-30%. We observe a similar trend for angle prediction: machine learning approaches can achieve 30-50% improvement upon the baseline model for less than 12-hour lead time predictions and 10-20% improvement for longer lead time predictions.

In addition, we observe that the ensemble model outperforms the best single machine learning model consistently across tasks and error measures. The advantage of an ensemble model is especially strong for less than 12-hour lead time predictions. The ensemble model can achieve 0-8% MAE reduction depending on the specific lead time (except for a slightly worse performance on the longer-term expected shortfall for speed).

Lead Time	Metric	Baseline	Elastic Net	Decision Tree	Random Forest	LightGBM	XGBoost	Ensemble
1		0.96	0.54	0.56	0.53	0.49	0.49	**0.48**
2		1.05	0.71	0.76	0.71	0.64	0.65	**0.63**
3		1.18	0.80	0.85	0.8	0.72	0.72	**0.70**
6	MAE	1.61	0.91	0.97	0.91	0.85	0.85	**0.84**
12	($m.s^{-1}$)	1.94	1.0	1.05	0.99	0.96	0.96	**0.94**
24		1.37	1.07	1.11	1.08	1.06	1.06	**1.04**
36		2.13	1.17	1.20	1.17	1.16	1.16	**1.15**
48		1.57	**1.17**	1.22	1.18	**1.17**	**1.17**	**1.17**
1		2.37	1.40	1.48	1.36	1.27	1.28	**1.26**
2		2.60	1.80	1.95	1.77	1.63	1.65	**1.61**
3	Expected	2.94	2.01	2.16	1.99	1.81	1.82	**1.78**
6	Shortfall	3.82	2.27	2.45	2.25	2.15	**2.14**	**2.14**
12	85%	4.55	2.48	2.64	2.44	**2.38**	2.40	2.39
24		3.51	2.61	2.77	2.60	**2.58**	**2.58**	2.59
36		4.93	2.79	2.89	2.76	**2.75**	**2.75**	2.77
48		4.03	2.80	2.95	2.79	**2.78**	**2.78**	2.81

Table 6: Beta test results on the test set for wind speed prediction for all models. We record the MAE and expected shortfall at 85% level for different lead times ranging from 1 hour to 48 hours.

Lead Time	Metric	Baseline	Elastic Net	Decision Tree	Random Forest	LightGBM	XGBoost	Ensemble
1		25	26	14	13	**12**	13	**12**
2		26	31	20	18	17	17	**16**
3		29	35	23	21	19	19	**18**
6	MAE	41	40	29	28	**24**	**24**	**24**
12	($m.s^{-1}$)	58	43	35	33	31	31	**30**
24		42	45	40	38	37	38	**36**
36		70	49	45	42	42	42	**41**
48		52	48	47	44	44	44	**42**
1		73	82	59	53	**51**	**51**	**51**
2		79	103	77	72	**68**	**68**	**68**
3	Expected	89	117	90	83	76	75	**74**
6	Shortfall	115	132	108	107	**94**	95	95
12	85%	136	138	127	125	**117**	**117**	**117**
24		**127**	143	138	135	134	135	132
36		155	148	145	**140**	**140**	**140**	**140**
48		145	148	146	**143**	**143**	**143**	**143**

Table 7: Beta test results on the test set for wind direction prediction for all models. We record the MAE and expected shortfall at 85% level for different lead times ranging from 1 hour to 48 hours.

3.2. Prescriptive Methodology Results

Table 8 compares the performance of several models for recommending binary hourly actions (anticipating dangerous conditions or maintaining production levels). It includes the baseline model, the previous Elastic Net ensemble model, and a series of Optimal Policy Trees (OPT) with different health costs associated with false negatives (-4000, -6000, -10000, -15000, and -20000). Recall that the health cost used to train OPTs is a parameter to tune conservatism towards pollution of our models: a higher cost leads to more cautious care towards recommending operation at normal levels. Table 8 reports the number of false positives and false negatives for each model. A false positive refers to when the plant operates with reduced production levels and undertakes actions to mitigate the odor impact, while in reality, these actions are unnecessary. A false positive is therefore associated with a loss of $2000 corresponding to the anti-odor injection costs and consequences of reduced production. On the other hand, a false negative refers to when the plant is operating at normal levels, while in reality, weather conditions are unfavorable, and pollution is carried to the city, leading to an air pollution incident.

As a remark, since the implementation of this component is underway, and we do not track the actual decisions undertaken by operators, we showcase back-testing results using

the baseline model as a benchmark. We simulate decisions using forecasts from the baseline and ensemble models and translate them into decisions using Table 1. The actual decisions can deviate from the simulated decisions because operators make decisions based on forecasts and expertise.

In general, our framework can lead to reductions in both false positives and false negatives. Specifically, looking at Optimal Policy Tree models, different choices of health cost lead to different levels of conservatism, which gives the modeler the space to explore the trade-off between cost savings and pollution mitigation goals. As the health cost increases, false negatives decrease, and false positives increase. Comparing the OPT models with the baseline model shows that the OPT models have overall better performance, as they have lower health costs and fewer false positives.

Model	**Health Cost**	**False Positives**	**False Negatives**	**Cost savings**	**Pollution Reduction**
Baseline		288	110	0%	0%
Ensemble		51	133	82%	-21%
Optimal Policy Tree	-4000	20	113	93%	-3%
	-6000	32	102	89%	7%
	-10000	106	74	63%	33%
	-15000	174	58	40%	47%
	-20000	282	38	2%	65%

Table 8: Performance of three families of models for recommending actions. We include the baseline model, the Elastic Net ensemble, and a series of optimal policy trees trained with different health costs associated with false negatives.

In addition, the OPTs provide interpretable insights on how the different ensemble members are used to prescribe, as illustrated by Figure 4 below. In particular, we notice that a simple tree like the one corresponding to choosing a health cost of $15,000 (tree on the right below) can reduce pollution emissions during dangerous scenarios by 47% and save 40% of the unnecessary costs. Conveniently, it also relies on only three ensemble members: XGBoost and Elastic Net predicting speed, and Random Forest predicting the cosine component of the wind direction. It also suggests that different ensemble members capture different aspects of the data and together make better recommendations.

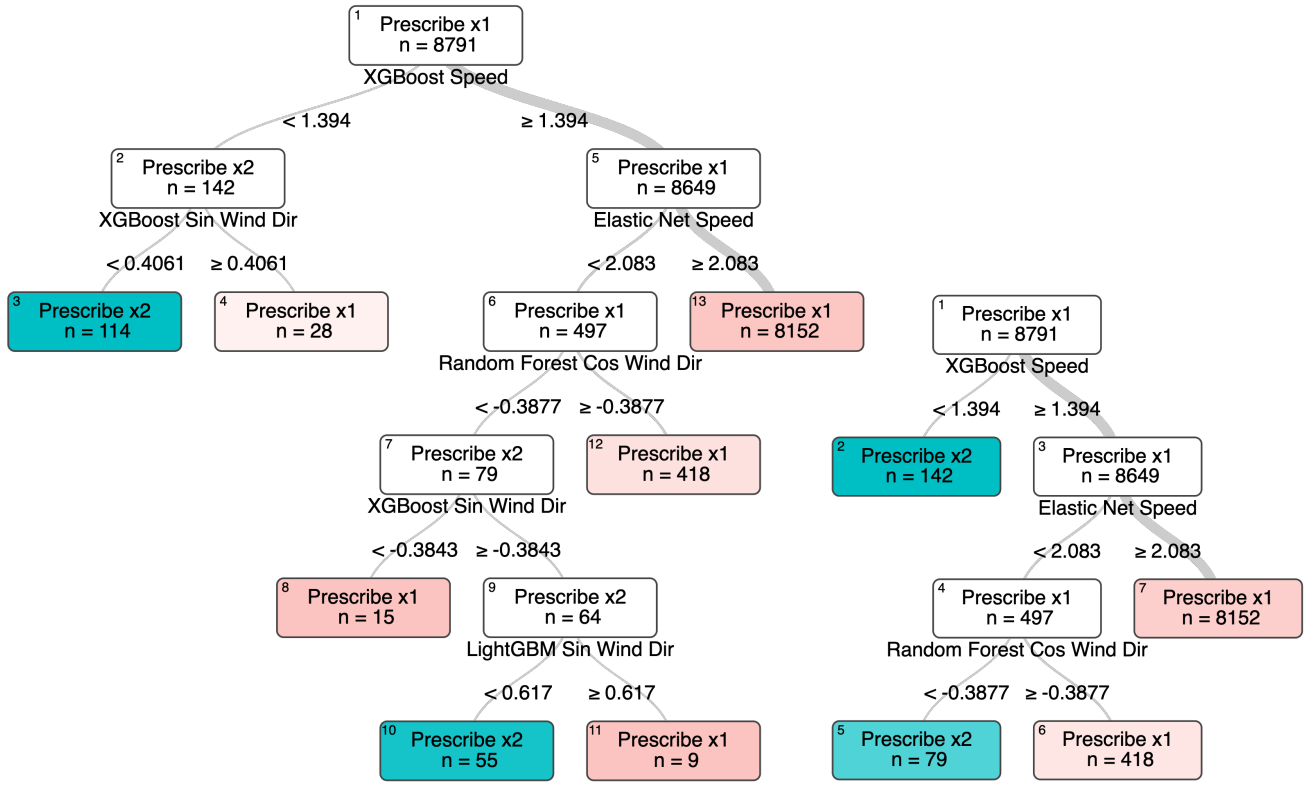

Figure 4: Optimal policy trees trained using a health cost of \$10,000 (left) and \$15,000 (right). The trees illustrate an interpretable decision-making process to arrive at certain recommended decisions (prescription options). `Prescribe x1` corresponds to maintaining production rate while `Prescribe x2` corresponds to reducing plant operations and injecting odor control chemicals.

4. Conclusion

In conclusion, our study introduces a novel and data-driven solution to mitigate the harmful effects of air pollution caused by industrial plants in urban areas. We provide a comprehensive solution for managing industrial operations and weather-related risks by combining advanced weather forecasting and decision-making models. Our framework, which incorporates both predictive and prescriptive machine learning models, was successfully implemented at the OCP Safi production site, resulting in improved forecasting accuracy and decision-making efficiency. Given the crucial role of weather in industrial environmental impact, we believe that our approach can be adapted and effectively applied in similar settings.

Our framework has demonstrated its value in managing air pollution in chemical production sites, and the results achieved at the OCP Safi site hold the potential to inspire a more sustainable and responsible chemical production industry globally. The flexibility and adaptability of our approach enable its core components of data enhancement, real-time monitoring, and prescriptive models to be universally applied to different chemical factories. Although each production site presents unique challenges, our data-driven approach can be customized to meet the needs and conditions of each location. Utilizing the latest advancements in weather forecasting and data analysis, we aim to assist factories in effectively managing air pollution and promoting the safety and well-being of the surrounding communities.

References

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, E. Bezirtzoglou, Environmental and health impacts of air pollution: a review, Frontiers in public health (2020) 14.

[2] B. Brunekreef, S. T. Holgate, Air pollution and health, The lancet 360 (9341) (2002) 1233–1242.

[3] A. L. Power, R. K. Tennant, R. T. Jones, Y. Tang, J. Du, A. T. Worsley, J. Love, Monitoring impacts of urbanisation and industrialisation on air quality in the anthropocene using urban pond sediments, Frontiers in Earth Science 6 (2018) 131.

[4] Z. Tong, K. M. Zhang, The near-source impacts of diesel backup generators in urban environments, Atmospheric Environment 109 (2015) 262–271. doi:https://doi.org/10.1016/j.atmosenv.2015.03.020.
URL https://www.sciencedirect.com/science/article/pii/S1352231015002381

[5] Z. Tong, B. Yang, P. K. Hopke, K. M. Zhang, Microenvironmental air quality impact of a commercial-scale biomass heating system, Environmental Pollution 220 (2017) 1112–1120. doi:10.1016/j.envpol.2016.11.025.
URL https://www.sciencedirect.com/science/article/pii/S0269749116321492

[6] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, A. Osornio-Vargas, A systematic review of data mining and machine learning for air pollution epidemiology, BMC Public Health 17 (1) (2017) 907. doi:10.1186/s12889-017-4914-3.

[7] T. Madan, S. Sagar, D. Virmani, Air Quality Prediction using Machine Learning Algorithms –A Re-

view, in: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 140–145. doi:10.1109/ICACCCN51052.2020.9362912.

[8] Doreswamy, H. K s, Y. Km, I. Gad, Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models, Procedia Computer Science 171 (2020) 2057–2066. doi:10.1016/j.procs.2020.04.221.
URL https://www.sciencedirect.com/science/article/pii/S1877050920312060

[9] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, L. Vanneschi, A Machine Learning Approach to Predict Air Quality in California, Complexity 2020 (2020) e8049504, publisher: Hindawi. doi:10.1155/2020/8049504.
URL https://www.hindawi.com/journals/complexity/2020/8049504/

[10] D. Sanjeev, Implementation of Machine Learning Algorithms for Analysis and Prediction of Air Quality, International Journal of Engineering Research 10 (03).

[11] K. Kumar, B. P. Pande, Air pollution prediction with machine learning: a case study of Indian cities, International Journal of Environmental Science and Technology (May 2022). doi:10.1007/s13762-022-04241-5.
URL https://doi.org/10.1007/s13762-022-04241-5

[12] J. Zhu, Q. Zeng, A mathematical formulation for optimal control of air pollution, Science in China Series D: Earth Sciences 46 (10) (2003) 994–1002. doi:10.1007/BF02959394.
URL https://doi.org/10.1007/BF02959394

[13] Y. Panagopoulos, C. Makropoulos, M. Mimikou, Decision support for diffuse pollution management, Environmental Modelling & Software 30 (2012) 57–70. doi:10.1016/j.envsoft.2011.11.006.

[14] SwissAid, Négociants suisses et engrais dangereux : violations de droits humains au maroc (2018).
URL https://voir-et-agir.ch/content/uploads/2018/12/Rapport_Maroc.pdf

[15] S. Grange, Technical note: Averaging wind speeds and directions (06 2014). doi:10.13140/RG.2.1.3349.2006.

[16] D. H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259. doi:https://doi.org/10.1016/S0893-6080(05)80023-1.

[17] M. Amram, J. Dunn, Y. D. Zhuo, Optimal policy trees, Machine Learning 111 (7) (2022) 2741–2768. doi:10.1007/s10994-022-06128-5.
URL https://doi.org/10.1007/s10994-022-06128-5

[18] G. Van Rossum, F. L. Drake Jr, Python tutorial, Centrum voor Wiskunde en Informatica Amsterdam,

The Netherlands, 1995.

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

[20] L. Interpretable AI, Interpretable ai documentation (2023).
URL `https://www.interpretable.ai`