Supplementary Materials

A Novel Explainable Deep Learning Method Based on Spectral Co-Clustering
for Ozone Time Series Forecasting

Introduction

This document provides additional explainability results uusing spectral clustering results and analyzes as a supplement to the main paper.

1. Extending Section 5.6.1: Spectral Coclustering Details for All Stations

To enhance the interpretability of the forecasting models, the Spectral Co-Clustering method is applied to identify valuable subgroups within each station's data. With this approach, instances and features with similar forecasting patterns can be clustered together, offering additional information on why certain models are more accurate. To determine the optimal number of clusters for each station's data, the elbow method is employed so that the split represents the highest-value variations in ozone dynamics. With this co-cluster analysis, the predictive power of the models can be better understood, and the hidden patterns and dependencies within the co-clusters can be revealed.

1.1. Forecasting Performance Analysis Across Co-Clusters

After clustering each station's data set using spectral co-clustering, the prediction performance of the best model is evaluated using all co-clusters of each station's data. This analysis helps determine whether certain clusters contain patterns that enhance prediction accuracy or, conversely, introduce complexity that degrades performance. For each station, the RMSE, MSE, and MAE values are computed across co-clusters to observe how the forecasting accuracy varies. The co-clusters on which the best models are achieved with lower RMSE values contain instances with high temporal dependencies and highly correlated features. Conversely, higher RMSE co-clusters usually correspond to irregular periods of ozone variations, external environmental influences, or weak feature correlation. These insights help to understand model performance and potential areas for improvement. The optimal hyperparameters achieved by the Bayesian optimization algorithm for the best and worst co-clusters evaluation using the best models for each dataset are provided in Table 1. The co-clusters are represented in the form of (x,y), where x denotes groups of data instances (e.g., time intervals) and y denotes groups of features contributing to ozone variation. The number of optimal co-clusters differs across stations because the elbow technique identifies the optimal cluster number based on the natural ozone variation patterns in the individual datasets. These patterns are influenced by local environmental and meteorological conditions and, therefore, assume different optimal cluster numbers even when the number of the data instances are the same.

The co-clustering analysis for the Aljarafe dataset using the best-performing model (HMAM) provides valuable insights into how different subsets of data impact forecasting performance (Table 2). Among 36 co-clusters, Cluster (1,1) demonstrates the best forecasting performance, achieving an RMSE of 6.02 and MAE of 4.13, significantly lower than most other clusters. This suggests that the instances and features grouped within this cluster exhibit strong predictive relationships, allowing the model to capture ozone fluctuations effectively. Conversely, Cluster (2,4) exhibits the highest RMSE (26.18) among all clusters, indicating poor forecasting performance. The high MAE of 20.93 suggests large deviations in predictions, likely due to erratic ozone variations or weak feature correlations. Furthermore, a general pattern observed across the co-clusters is that those containing more informative feature subsets (e.g., Clusters with 8 and 9 features) tend to achieve lower RMSE values, while those with fewer features (e.g., Clusters with only 1 feature) result in significantly degraded performance. This highlights the importance of feature selection and

Table 1: Optimal hyperparameters achieved by Bayesian optimization for the best and worst co-clusters of each dataset.

Cluster	Metric	Aljarafe	Asomadilla	Bermejales	Ronda del Valle	Torneo
Best	Cluster ID	(1,1)	(1,3)	(3,1)	(0,6)	(4,0)
	Activation	ReLU	Tanh	ReLU	Sigmoid	Sigmoid
	Batch Size	41	16	47	16	21
	Dropout Rate	0.43	0.16	0.12	0.17	0.15
	Epochs	78	74	87	99	96
	Num Units	58	94	124	51	105
Worst	Cluster ID	(2,4)	(3,4)	(1,5)	(0,3)	(0,4)
	Activation	ReLU	Tanh	Tanh	Tanh	Tanh
	Batch Size	61	16	20	39	25
	Dropout Rate	0.24	0.21	0.11	0.28	0.10
	Epochs	80	70	85	96	62
	Num Units	64	99	97	127	102

clustering in enhancing forecasting accuracy. Additionally, co-clusters with larger instance counts (such as Clusters in the 1st row) tend to show better generalization, whereas smaller clusters (e.g., Cluster (4,4) with only 4374 instances) suffer from data sparsity, leading to higher prediction errors as shown in Table 2.

Table 2: Summary of Co-Cluster Analysis for Aljarafe Dataset. This table presents the details of error metrics (MSE, MAE, RMSE) for multiple co-clusters within the Aljarafe dataset with best model Hybrid Model with Attention Mechanism (HMAM), including the number of instances and features for each cluster.

Cluster ID	RMSE	MSE	MAE	Instances	Features	Cluster ID	RMSE	MSE	MAE	Instances	Features
(0, 0)	14.13	199.65	11.07	8292	6	(3, 0)	15.52	240.87	12.21	7187	6
(0, 1)	7.14	50.97	5.23	8292	8	(3, 1)	7.41	54.90	5.30	7187	8
(0, 2)	16.56	275.72	13.04	8292	7	(3, 2)	17.58	309.05	14.13	7187	7
(0, 3)	13.09	171.34	10.23	8292	9	(3, 3)	12.65	160.02	9.67	7187	9
(0, 4)	18.39	338.19	14.62	8292	1	(3, 4)	16.31	266.01	12.79	7187	1
(0, 5)	14.52	210.83	11.52	8292	1	(3, 5)	15.18	230.43	12.18	7187	1
(1, 0)	16.20	262.44	12.11	9662	6	(4, 0)	18.21	331.60	14.90	4374	6
(1, 1)	6.02	36.24	4.13	9662	8	(4, 1)	8.33	68.38	6.12	4374	8
(1, 2)	18.07	326.52	13.96	9662	7	(4, 2)	18.64	347.44	15.01	4374	7
(1, 3)	10.91	119.02	8.47	9662	9	(4, 3)	13.15	172.92	10.14	4374	9
(1, 4)	22.82	520.75	18.00	9662	1	(4, 4)	20.16	406.42	16.57	4374	1
(1, 5)	16.37	267.97	12.74	9662	1	(4, 5)	17.60	309.76	14.25	4374	1
(2, 0)	23.10	533.61	18.36	7608	6	(5,0)	17.15	294.12	13.41	7762	6
(2, 1)	7.87	61.93	5.39	7608	8	(5,1)	8.47	71.74	6.12	7762	8
(2, 2)	20.09	403.60	15.62	7608	7	(5,2)	17.33	300.32	13.82	7762	7
(2, 3)	17.89	320.05	13.58	7608	9	(5,3)	13.35	178.22	9.98	7762	9
(2, 4)	26.18	685.39	20.93	7608	1	(5,4)	20.13	405.21	16.01	7762	1
(2, 5)	20.49	419.84	16.41	7608	1	(5,5)	16.91	285.94	13.53	7762	1
Best Cluster (1, 1)	6.02	36.24	4.13	9662	8	Worst Cluster (2, 4)	26.18	685.39	20.93	7608	1

The variations in co-clustering outcomes for the Asomadilla dataset using the best model (HMWF) highlight the impact of instance distributions and feature composition as shown in Table 3. Among the 25 co-clusters, cluster (1,3) achieved the lowest RMSE (8.89) and MAE (7.46), indicating better forecasting performance. The cluster consists of 320 instances and 4 features, which suggests that a few but relevant features are involved in improved predictive performance. The feature set in this cluster enhances the model ability to capture the critical temporal relationships of ozone variation and, thereby, reduces the prediction error. Moreover, when compared with other clusters with similar numbers of instances but different feature sets, such as (1,1) and (1,2), which have greater RMSE values (12.07 and 12.74, respectively), it shows the importance of feature selection in reducing forecasting error. In contrast, the cluster (3,4) has the worst predictive accuracy with the highest RMSE (20.93) and MAE (17.75). The cluster contains 281 samples but lacks any informative features (0 features selected) that would impact the predictive strength of the model. The absence of key temporal or pollutant-related features prevents the model from effectively learning

patterns, resulting in substantial errors. Additionally, clusters with similar instance counts but more features, such as (3,1) and (3,2), have relatively lower RMSE values (20.09 and 20.06, respectively), which shows that even a few good features can improve the forecasting performance greatly.

In short, the Asomadilla dataset results show that clusters with 4 selected features have better performance than clusters with smaller or no features (Table 3). For instance, clusters (0,3), (1,3), and (2,3) have consistently lower values of RMSE (13.50, 8.89, and 13.70, respectively), whereas clusters that contain few or no features, such as (0,4), (3,4), and (4,4), have significantly larger values of RMSE. This confirms the importance of proper feature selection in forecasting time series. Also, the clusters with low instance numbers (e.g., row 4 clusters) are discovered to have higher errors, perhaps due to insufficient instances for learning temporal patterns precisely. For example, cluster (4,0) with only 154 instances has a significantly higher RMSE of 15.78, whereas cluster (1,0) with 320 instances achieves a much lower RMSE of 9.71 with the same number of selected features (2). This suggests that feature selection is valuable but that the number of observations also has a serious impact on forecasting performance.

Table 3: Summary of Co-Cluster Analysis for Asomadilla Dataset. This table presents the details of error metrics (MSE, MAE, RMSE) for multiple co-clusters within the Asomadilla dataset with best model Hybrid Model with Weighted Fusion (HMWF), including the number of instances and features for each cluster.

Cluster ID	RMSE	MSE	MAE	Instances	Features	Cluster ID	RMSE	MSE	MAE	Instances	Features
(0, 0)	15.85	251.22	12.05	257	2	(3, 0)	16.98	288.32	13.83	281	2
(0, 1)	15.91	253.12	12.33	257	6	(3, 1)	20.09	403.60	16.96	281	6
(0, 2)	15.86	251.53	12.30	257	1	(3, 2)	20.06	402.40	16.72	281	1
(0, 3)	13.50	182.25	10.63	257	4	(3, 3)	13.12	172.13	10.74	281	4
(0, 4)	15.42	137.77	12.04	257	0	(3, 4)	20.93	438.06	17.75	281	0
(1, 0)	9.71	94.28	7.48	320	2	(4, 0)	15.78	249.00	13.18	154	2
(1, 1)	12.07	145.68	10.15	320	6	(4, 1)	16.65	277.22	14.47	154	6
(1, 2)	12.74	162.30	10.88	320	1	(4, 2)	16.30	265.69	13.78	154	1
(1, 3)	8.89	79.03	7.46	320	4	(4, 3)	12.26	150.30	10.12	154	4
(1, 4)	13.29	176.62	11.13	320	0	(4, 4)	16.82	282.91	14.86	154	0
(2, 0)	12.53	157.00	10.81	267	2						
(2, 1)	14.55	211.70	12.48	267	6						
(2, 2)	14.95	223.50	12.50	267	1						
2, 3)	13.70	187.69	11.72	267	4						
(2, 4)	15.22	231.64	12.76	267	0						
Best Cluster (1, 3)	8.89	79.03	7.46	320	4	Worst Cluster (3, 4)	20.93	438.06	17.75	281	0

In the case of the Bermejales dataset, the smallest RMSE (5.79) occurs in a cluster (3,1) with 5 features, followed by clusters (4,1) (RMSE = 6.20) and (2,1) (RMSE = 6.26), which means that this set of features contributes to better prediction accuracy. The largest RMSE (20.15) for a cluster (1,5) containing 7 features suggests that the presence of many features introduces noise into the model and makes it perform worse. The overall trend reveals that the clusters with 5 features always perform better, and the ones with 7 or more features (for example, (1,5), (5,5), and (4,5)) contain higher error rates, thus providing more support to optimal feature selection. Moreover, the larger instance clusters such as (0,0) (2787 instances) and (1,0) (3307 instances) contain higher RMSE values (12.65 and 16.21, respectively), meaning that large sample sizes introduce more variability of predictions. Although for smaller instance sets such as (2,1) (1130 instances) and (4,1) (1719 instances), better RMSE is obtained, this once again shows the significance of well-distributed instances. Interestingly, cluster (3,1) performs the best among all, demonstrating that proper choice of feature subset with an average number of instances is crucial to obtain the best forecasting performance. Overall, the results emphasize that correct feature selection and instance clustering with an appropriate balance between instances can significantly enhance predictive performance, while redundant features and large sample variations negatively impact performance.

Moreover, co-cluster analysis of Ronda Del Valle data reveals that clusters with 10 features such as (0,6) (RMSE = 8.62), (1,6) (RMSE = 9.75), (2,6) (RMSE = 10.51), and (6,6) (RMSE = 9.33) have the least RMSE values, indicating the performance is better for more number of features for this data (Table 5). Conversely, single-feature clusters such as (0,3) (RMSE = 30.65), (3,3) (RMSE = 28.82), and (5,3) (RMSE = 28.16) exhibit the highest RMSE, which displays that with a minimum set of features, low-quality forecasting

Table 4: Summary of Co-Cluster Analysis for Bermejales Dataset. This table presents the details of error metrics (MSE, MAE, RMSE) for multiple co-clusters within the Bermejales dataset with best model HMWF, including the number of instances and features for each cluster.

Cluster ID	RMSE	MSE	MAE	Instances	Features	Cluster ID	RMSE	MSE	MAE	Instances	Features
(0, 0)	12.65	160.02	9.94	2787	9	(3, 0)	11.05	122.10	7.60	2919	9
(0, 1)	7.74	59.90	5.61	2787	5	(3, 1)	5.79	33.52	3.36	2919	5
(0, 2)	13.69	187.41	10.44	2787	2	(3, 2)	10.17	103.42	7.05	2919	2
(0, 3)	11.87	140.89	9.57	2787	7	(3, 3)	12.76	162.81	9.93	2919	7
(0, 4)	15.24	232.25	12.30	2787	3	(3, 4)	11.93	142.32	8.92	2919	3
(0, 5)	15.82	250.27	12.77	2787	7	(3, 5)	12.35	152.52	8.93	2919	7
(1, 0)	16.21	262.76	13.00	3307	9	(4, 0)	13.65	186.32	10.68	1719	9
(1, 1)	8.24	67.89	5.92	3307	5	(4, 1)	6.20	38.44	4.40	1719	5
(1, 2)	18.18	330.51	14.25	3307	2	(4, 2)	16.64	276.88	12.57	1719	2
(1, 3)	15.86	251.53	12.43	3307	7	(4, 3)	18.14	329.05	14.28	1719	7
(1, 4)	17.28	298.59	13.59	3307	3	(4, 4)	15.37	236.23	11.55	1719	3
(1, 5)	20.15	406.02	15.74	3307	7	(4, 5)	18.15	329.42	13.75	1719	7
(2, 0)	17.81	317.19	13.86	1130	9	(5,0)	15.60	243.36	12.50	2623	9
(2, 1)	6.26	39.18	4.43	1130	5	(5,1)	8.09	65.44	6.01	2623	5
(2, 2)	18.70	349.69	14.82	1130	2	(5,2)	18.90	357.21	15.26	2623	2
(2, 3)	15.23	231.95	12.12	1130	7	(5,3)	15.42	237.77	12.48	2623	7
(2, 4)	15.36	235.92	11.89	1130	3	(5,4)	16.64	276.88	13.06	2623	3
(2, 5)	17.11	292.75	13.93	1130	7	(5,5)	18.33	335.98	14.85	2623	7
Best Cluster (3, 1)	5.79	33.52	3.36	9662	5	Worst Cluster (1, 5)	20.15	406.02	15.72	7608	1

performance is obtained. Concerning the instance size, the larger-sized clusters (4,0) (16,570 instances, RMSE = 18.73) and (6,0) (15,379 instances, RMSE = 19.09) perform relatively well, and the smaller clusters such as (3,3) (4,653 instances, RMSE = 28.82) achieved much larger RMSE values, which indicate that better model stability arises with higher sample sizes. However, several moderate-sized clusters, such as (1,1) (5,504 instances, RMSE = 12.19) and (6,6) (15,379 instances, RMSE = 9.33), are even lower in RMSE, confirming that the best feature and instance combinations are responsible for accuracy in this dataset. Furthermore, a comparison of feature subsets indicates that groups of five features, such as (1,0) (RMSE = 18.42) and (2,0) (RMSE = 19.89), generally have lower RMSE than seven-feature groups, such as (2,2) (RMSE = 20.67) and (5,2) (RMSE = 23.03), demonstrating that there is a point at which too many features can induce redundancy and degrade performance. Two-feature clusters tend to have moderate performance, with RMSE between 12.19 and 24.47, which means that although fewer features might be useful in certain situations, they can also result in poor generalization. The findings emphasize the need to balance instance distribution and feature selection to achieve optimal performance using five to ten features. Medium to large instance sizes enhance model stability, while small clusters produce higher errors. The study indicates that co-clustering maximizes feature-instance pairings for better forecasting.

Finally, Table 6 presents a comprehensive analysis of error metrics (RMSE, MSE, MAE) across different co-clusters in the Torneo dataset using the HMAM model. The best-performing cluster (4,0) has the lowest RMSE (0.94), MSE (0.88), and MAE (0.63) with the highest frequency (14,281) and eight features. This shows that bigger data with a better-structured feature set help improve model generalization and predictability. Looking at different clusters, there is a general pattern: clusters with higher numbers of features have lower values of RMSE, though there are outliers. For instance, cluster (3,2) has an RMSE of 1.94, while (3,0) has an RMSE of 9.67, though both have the same number of instances (7,835) with a different number of features, which means that the quality and number of the selected features matters. Similarly, clusters (1,3) and (4,3) have low RMSE scores (0.98 and 1.15, respectively), highlighting that certain feature subsets allow the model to perform exceptionally well. One-feature clusters, such as (0,2) and (3,2), acted very differently with RMSEs of 10.77 and 1.94, respectively. This suggests that certain single features contribute significantly to prediction, whereas others have little useful information or even create noise. Furthermore, cluster (5,2) has an RMSE of 3.56 with a comparatively small number of instances (11,612), indicating that feature relevance is more significant than data volume. The general trend indicates that the use of a good mix of instances and features leads to the best model performance. Clusters with extremely high RMSE, i.e., (0,4), indicate

Table 5: Summary of Co-Cluster Analysis for Ronda Del Valle Dataset. This table presents the details of error metrics (MSE, MAE, RMSE) for multiple co-clusters within the Ronda Del Valle dataset with the best model HMWF, including the number of instances and features for each cluster.

Cluster ID	RMSE	MSE	MAE	Instances	Features	Cluster ID	RMSE	MSE	MAE	Instances	Features
(0, 0)	20.00	400.00	16.73	10251	5	(4, 0)	18.73	350.81	14.93	16570	5
(0, 1)	24.47	598.78	19.72	10251	2	(4, 1)	21.65	468.72	17.12	16570	2
(0, 2)	21.34	455.39	16.32	10251	7	(4, 2)	24.15	583.22	19.40	16570	7
(0, 3)	30.65	939.42	25.26	10251	1	(4, 3)	26.44	699.07	21.48	16570	1
(0, 4)	23.67	560.26	19.46	10251	4	(4, 4)	22.84	521.66	18.54	16570	4
(0, 5)	29.99	899.40	24.09	10251	1	(4, 5)	26.20	686.44	21.40	16570	1
(0, 6)	8.62	72.30	6.32	10251	10	(4, 6)	10.22	104.44	7.69	16570	10
(1, 0)	18.42	339.29	14.06	5504	5	(5, 0)	21.88	478.73	17.04	7071	5
(1, 1)	12.19	148.59	8.62	5504	2	(5, 1)	20.32	412.90	15.28	7071	2
(1, 2)	17.91	320.76	13.47	5504	7	(5, 2)	23.03	530.38	18.40	7071	7
(1, 3)	19.40	376.36	14.47	5504	1	(5, 3)	28.16	792.98	22.64	7071	1
(1, 4)	19.13	365.95	13.66	5504	4	(5, 4)	22.76	518.01	17.88	7071	4
(1, 5)	16.02	256.64	11.50	5504	1	(5, 5)	27.75	770.06	22.42	7071	1
(1, 6)	9.75	95.06	7.10	5504	10	(5, 6)	14.11	199.09	11.04	7071	10
(2, 0)	19.89	395.61	15.49	13607	5	(6,0)	19.09	364.42	14.32	15379	5
(2, 1)	18.80	353.44	15.21	13607	2	(6,1)	21.59	466.12	17.29	15379	2
(2, 2)	20.67	427.24	16.52	13607	7	(6,2)	19.04	362.52	14.50	15379	7
(2, 3)	23.24	530.38	18.59	13607	1	(6,3)	24.25	588.06	18.86	15379	1
(2, 4)	20.82	433.47	16.59	13607	4	(6,4)	21.68	470.02	17.07	15379	4
(2, 5)	23.03	530.38	18.31	13607	1	(6,5)	24.19	585.15	18.60	15379	1
(2, 6)	10.51	110.46	7.66	13607	10	(6,6)	9.33	87.04	6.42	15379	10
(3, 0)	20.19	407.63	15.62	4653	5						
(3, 1)	22.18	491.95	16.97	4653	2						
(3, 2)	23.89	570.73	19.24	4653	7						
(3, 3)	28.82	830.59	23.27	4653	1						
(3, 4)	26.21	686.96	21.07	4653	4						
(3, 5)	27.42	751.85	21.43	4653	1						
(3, 6)	14.68	215.50	12.27	4653	10						
Best Cluster (0, 6)	8.62	72.30	6.32	10251	10	Worst Cluster (0, 3)	30.65	939.42	25.26	10251	1

overfitting or weak feature representation, whereas clusters with low RMSE, i.e., (4,0) and (1,3), reflect the advantages of well-structured feature sets. These observations are very useful in adjusting feature selection and grouping operations to improve prediction performance.

Overall, the spectral co-clustering approach provides valuable interpretability by demonstrating how certain feature and time instance combinations contribute to model success or failure. The performance variations across co-clusters emphasize the need for targeted feature selection strategies and adaptive model tuning to handle challenging data partitions effectively. These findings not only enhance the explainability of the forecasting model but also pave the way for improved decision-making in air quality monitoring.

2. Extending Section 5.6.2. Interpreting OR Visualizing Co-Clustered Feature and Temporal Patterns with XDeepSpecT Heatmaps Details for All Stations

In this section, the best and worst-performing co-clusters are visualized using heatmaps and other interpretability plots to demonstrate how the proposed XDeepSpecT method enhances understanding of the interaction between input features and prediction performance. The visualizations include heatmaps to highlight structured feature interactions, trend diagrams to illustrate key feature variations within the best and worst co-clusters, and temporal ozone distribution plots by hour and month to capture daily and seasonal patterns. These insights enhance interpretability by linking spectral co-clustering results with ozone formation dynamics across all five datasets.

Table 6: Summary of Co-Cluster Analysis for Torneo Dataset. This table presents the details of error metrics (MSE, MAE, RMSE) for multiple co-clusters within the Torneo dataset with best model HMAM, including the number of instances and features for each cluster.

Cluster ID	RMSE	MSE	MAE	Instances	Features	Cluster ID	RMSE	MSE	MAE	Instances	Features
(0, 0)	10.71	114.70	7.68	11150	8	(3, 0)	9.67	93.50	7.18	7835	8
(0, 1)	10.70	114.49	7.51	11150	9	(3, 1)	3.09	9.54	2.47	7835	9
(0, 2)	10.77	115.99	7.57	11150	1	(3, 2)	1.94	3.76	1.42	7835	1
(0, 3)	5.20	27.04	3.65	11150	3	(3, 3)	6.85	46.92	5.20	7835	3
(0, 4)	11.18	124.99	10.22	11150	7	(3, 4)	3.28	10.75	2.81	7835	7
(0, 5)	11.12	123.65	7.88	11150	5	(3, 5)	4.77	22.75	3.37	7835	5
(1, 0)	3.29	10.82	3.12	11740	8	(4, 0)	0.94	0.88	0.63	14281	8
(1, 1)	3.12	9.73	2.45	11740	9	(4, 1)	3.34	11.15	3.09	14281	9
(1, 2)	1.58	2.49	1.34	11740	1	(4, 2)	1.49	2.22	0.93	14281	1
(1, 3)	0.98	0.96	0.75	11740	3	(4, 3)	1.15	1.32	0.95	14281	3
(1, 4)	1.18	1.39	0.90	11740	7	(4, 4)	1.89	3.57	1.55	14281	7
(1, 5)	2.11	4.45	1.89	11740	5	(4, 5)	1.23	1.51	0.90	14281	5
(2, 0)	2.98	8.88	2.57	9035	8	(5, 0)	2.15	4.62	1.79	11612	8
(2, 1)	1.85	3.42	1.54	9035	9	(5, 1)	2.64	6.96	2.11	11612	9
(2, 2)	2.48	6.15	2.15	9035	1	(5, 2)	3.56	12.67	3.25	11612	1
(2, 3)	1.80	3.24	1.58	9035	3	(5, 3)	2.44	5.95	2.09	11612	3
(2, 4)	1.15	1.32	0.88	9035	7	(5, 4)	1.79	3.20	1.48	11612	7
(2, 5)	1.40	1.96	1.10	9035	5	(5, 5)	1.56	2.43	1.26	11612	5
Best Cluster (4, 0)	0.94	0.88	0.63	14281	8	Worst Cluster (0, 4)	11.18	124.99	10.22	11150	7

2.1. Explainability Analysis of Best and Worst Co-Cluster of Aljarafe Dataset

The proposed XDeepSpecT explainability approach for the Aljarafe dataset revealed significant differences in the accuracy of forecasts between co-clusters. The data for both the best co-cluster and worst co-cluster include instances from 2015 to 2023.

The co-cluster with the best performance accurately captures the temporal dynamics of ozone formation, with high correlations between lagged ozone, meteorological factors, and seasonal patterns, as shown in Fig. 1. The heatmap of this co-cluster shows that recent ozone levels (Lag 1–Lag 3) have high predictive power, validating that short-term previous ozone levels have a strong influence on future patterns (Fig. 1 (a)). In addition, higher-order lags (Lag 22–Lag 24) retain some influence, but with little reduced strength, suggesting the necessity of incorporating both short- and long-term dependencies in ozone forecasting. Temperature (TMP) is also an important feature, exhibiting a strong positive correlation with ozone levels, as expected from atmospheric chemistry principles, where higher temperatures promote photochemical reactions. Wind speed (WS), while not as dominant, has some variations that are synchronized with ozone variability, indicating its secondary role in controlling ozone dispersion. The Pearson Correlation Coefficient Heatmap (PCCH) of the best co-cluster further validates the effectiveness of the Explainable Deep Learning Approach with Spectral Co-Clustering for Time Series (XDeepSpecT) heatmap by revealing high feature correlations, indicating strong interrelationships among the selected features (Fig. 1 (c)). This structured coherence likely enhances the model's ability to capture meaningful patterns, leading to improved ozone predictions.

The temporal analysis also verifies the efficiency of this cocluster. A seasonal study indicates that the concentrations of ozone are highest from June to September, along with increased solar radiation and warmer temperatures favorable for photochemical reactions (Fig. 1 (d)). Atmospheric stability during the summer months further improves ozone accumulation, which confirms the need to incorporate seasonal dependencies while preparing forecasting models. Secondly, the ozone concentration is at its highest between 15:00 and 20:00, a pattern that is consistent with photochemical ozonogenesis induced by ultraviolet (UV) radiation (Fig. 1 (e)). With increasing sunlight intensity, ozone grows, reaching peak values in the late afternoon due to accumulated photochemical reactions. This daytime behavior justifies the fact that the model is aptly representing the cycle of ozone generation.

In contrast, the worst-performing co-cluster lacks critical temporal and meteorological correlations needed for accurate ozone prediction as shown in the last two rows of Fig. 1. The heatmap analysis shows little variance in historical ozone levels, making it difficult for the model to identify meaningful trends (Fig. 1 (f)).

The features in this group do not have high associations with ozone changes, which limits their predictive utility. Notably, the lack of substantial ozone fluctuations among timestamps indicates that the chosen characteristics do not represent the intrinsic dynamic nature of ozone generation. Furthermore, several parameters, such as wind direction (ALJARAFE-DD-4T_IN), show oscillations that are not consistent with ozone trends, making them useless predictors. Furthermore, the PCCH of the worst performing co-cluster exhibits weak correlations, particularly with a lower correlation (0.62) between key features. This lack of strong relationships hinders the model's ability to learn effective patterns, resulting in poor ozone prediction performance.

These findings emphasize the importance of selecting the feature set with strong temporal dependencies, meteorological effects, and seasonal trends. The improved performance of the optimal co-cluster suggests that ozone prediction models must pay greater attention to features that capture the atmospheric condition accurately, ensuring both prediction quality and interpretability. The spectral co-clustering approach can effectively cluster relevant data patterns well, enhancing explainability of deep learning models in air pollution forecasting.

2.2. Explainability Analysis of Best and Worst Co-Cluster of Asomadilla Dataset

Spectral co-clustering analysis using the XDeepSpecT approach of ozone forecasting at Asomodilla station highlights the contrast between the best and worst co-clusters in predictive accuracy and feature importance (Fig. 2). The data for both the best co-cluster and worst co-cluster include the instances from 2015.

The best-performing co-cluster effectively integrates significant meteorological and temporal features, i.e., temperature, wind speed, rolling mean of ozone concentration, and month, with the target ozone variable. The heatmap visualization of this co-cluster reveals evident patterns where temperature and wind speed have evident correlations with ozone levels (Fig. 2 (a)). These variables are significant in the capture of ozone dynamics as evidenced by their high relation with temporal fluctuations of ozone. The inclusion of the rolling mean serves to incorporate recent ozone variability, improving the model's ability to capture short-term variability. The inclusion of the month feature allows the model to understand seasonal trends in ozone, which is key for long-term forecast accuracy. Moreover, the PCCH shows a range of correlations between features, indicating important relationships between variables that enhance ozone prediction. Strong positive and negative correlations suggest that the model can use such dependencies for better forecasting (Fig. 2 (c)).

Temporal analysis confirms the suitability of the best co-cluster to explain ozone variation. Patterns on a monthly scale show an increasing trend in an upward direction from January to February in the concentration of ozone indicating probable seasonality and meteorological variability due to increased solar irradiation and stable atmospheric conditions. Patterns on an hourly scale show concentrations of ozone at their peak at the close of the afternoon, from 15:00 to 20:00, after the typical diurnal cycle of ozone production.

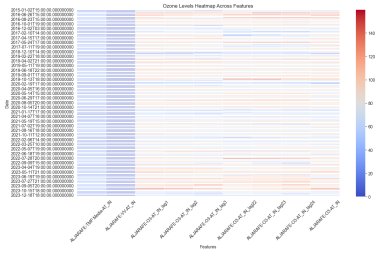
However, the low-performing co-cluster lacks informative features, which causes the inability to identify meaningful ozone trends. The poorly performing co-cluster heatmap is uniform, which indicates no useful patterns, significantly preventing the model from learning useful relationships (Fig. 2 (f)). The PCCH is displaying one characteristic with correlation 1, indicating no meaningful relationships between variables. Lack of interactions of various types does not allow the model to identify patterns and leads to poor performance in ozone prediction (Fig. 2 (h)). Without key meteorological and temporal variables, the model fails to account for essential factors influencing ozone formation and dispersion leading to worse performance on this cluster.

The comparative analysis of these co-clusters underscores the importance of carefully selecting features that reflect atmospheric and seasonal influences on ozone dynamics. The superior performance of the best co-cluster at Asomodilla station demonstrates that integrating temperature, wind velocity, rolling mean, and temporal features enhances the interpretability and predictive accuracy of deep learning-based ozone forecasting models.

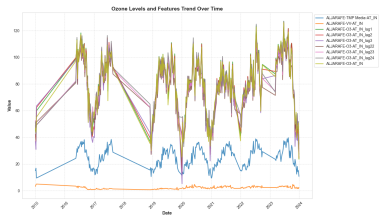
2.3. Explainability Analysis of Best and Worst Co-Cluster of Bermejales Dataset

The analysis of the best and worst co-clusters in the Bermejales dataset produces significant differences between the temporal correlations of ozone concentration and its corresponding features (Fig. 3).

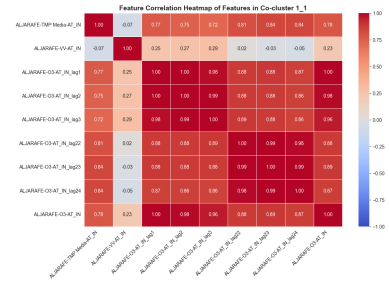
For the best co-cluster, the included features are lag 13 to lag 17, indicating that ozone levels at these previous time steps have the highest predictive power for the selected instances (Fig. 3 (a)). The trend plot of the time series confirms that the selected lags have a very strong correlation with the time-varying ozone levels (Fig. 3 (b)). The PCCH of the best co-cluster confirms the effectiveness of XDeepSpecT, showing



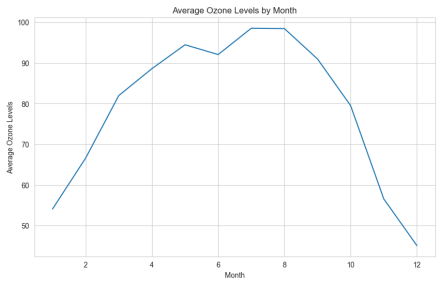
(a) Best co-cluster heatmap



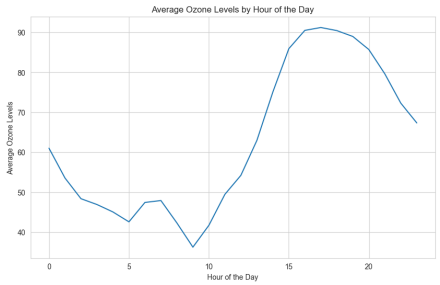
(b) Best co-cluster feature trends



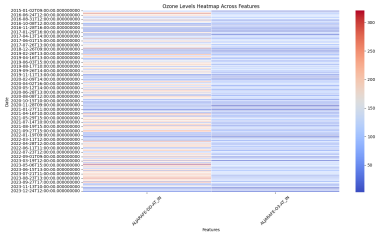
(c) PCCH of features in the best co-cluster



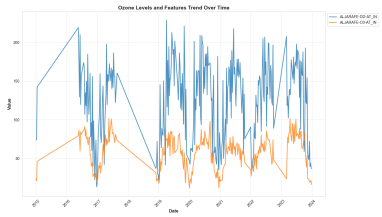
(d) Best co-cluster ozone levels by months



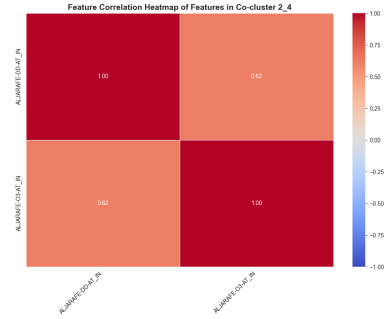
(e) Best co-cluster ozone level by hour of the day



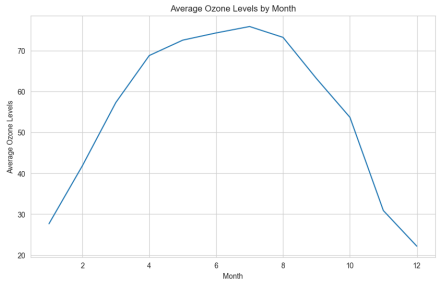
(f) Worst co-cluster heatmap



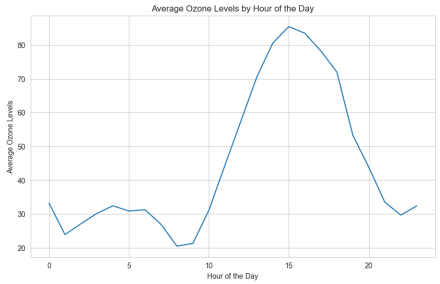
(g) Worst co-cluster feature trends



(h) PCCH of features in the worst co-cluster

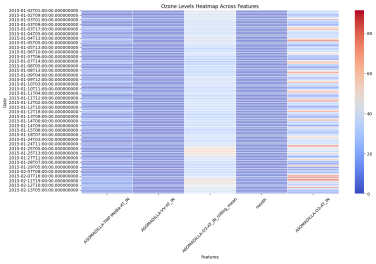


(i) Worst co-cluster ozone levels by months

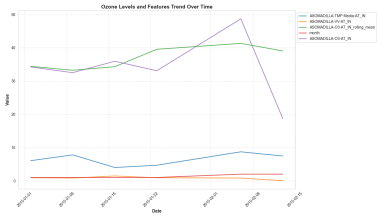


(j) Worst co-cluster ozone level by hour of the day

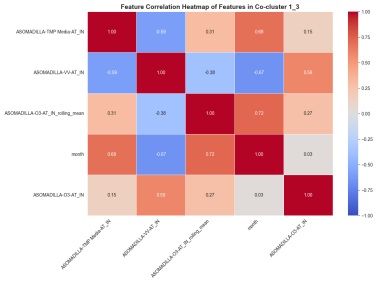
Figure 1: Analysis of the Best-Performing and Worst-Performing Co-Clusters in the Aljarafe Dataset: XDeepSpecT Heatmap, PCCH, Feature Trends, Hourly Variations, and Seasonal Patterns.



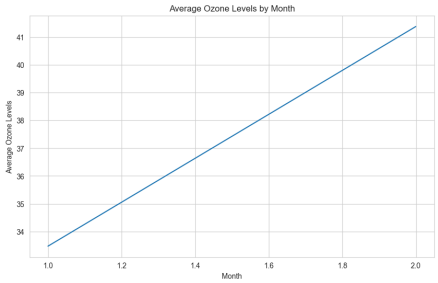
(a) Best co-cluster heatmap



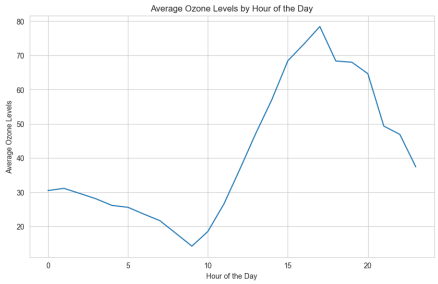
(b) Best co-cluster feature trends



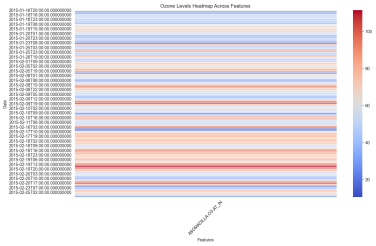
(c) Correlation heatmap features in best co-cluster



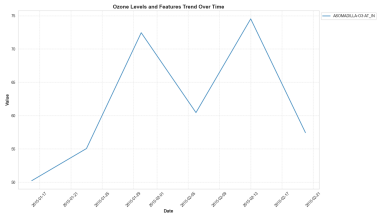
(d) Best co-cluster ozone levels by months



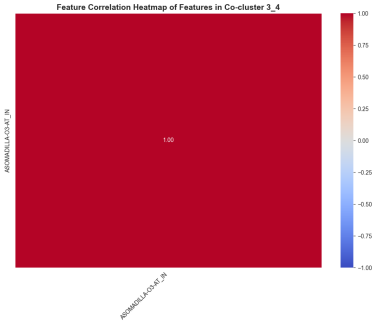
(e) Best co-cluster ozone level by hour of the day



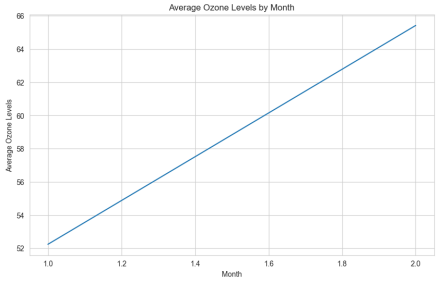
(f) Worst co-cluster heatmap



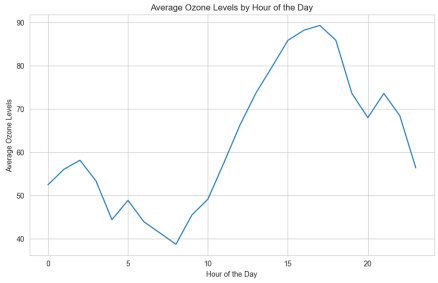
(g) Worst co-cluster feature trends



(h) Correlation heatmap features in worst co-cluster



(i) Worst co-cluster ozone levels by months



(j) Worst co-cluster ozone level by hour of the day

Figure 2: Analysis of the Best-Performing and Worst-Performing Co-Clusters in the Asomadilla Dataset: XDeepSpecT Heatmap, PCCH, Feature Trends, Hourly Variations, and Seasonal Patterns.

strong correlations among lag features (mostly above 0.85) (Fig. 3 (c)). This indicates a stable and coherent structure, making it well-suited for forecasting. The data for this co-cluster include instances from 2022 and 2023 with a clear ozone peak during July and August (months 7 and 8) (Fig. 3 (d)). This means that mid-to-late summer contains the most dense ozone, likely due to greater solar radiation and warmer temperatures, which accelerate photochemical reactions to create ozone. The respective heatmap shows evident differences between the selected features, indicating that ozone levels have consistent patterns that are well-captured by this feature set. The best co-cluster's hourly change of ozone graph displays a steep rise in ozone level during the day, peaking between 14:00 and 20:00 but with an unexpected drop around 18:00, before gradually falling off in the evening (Fig. 3 (e)). This pattern corresponds with typical ozone behavior, in the sense that maximum solar intensity power creates photochemical ozone, whereas after the sun has set the reaction rate drops and the concentration decreases.

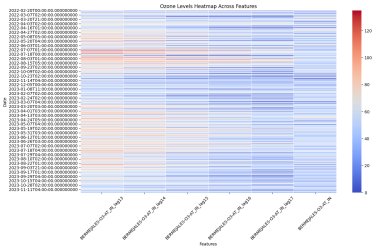
The worst co-cluster XDeepSpecT heatmap, on the other hand, includes lag features from lag 18 through lag 24, which implies that using longer-range dependencies in this case does not create strong prediction signals (Fig. 3 (f)). The PCCH of the worst co-cluster shows a decline in correlation over time, with values dropping below 0.60 for the latest lags (Fig. 3 (h)). This indicates increased variability and noise, reducing its reliability for prediction. The data used for this cluster are also in 2022 and 2023, and the ozone peaks are spread throughout April, May, July, and August (months 4, 5, 7, and 8) (Fig. 3 (i)). This indicates that ozone levels in this cluster do not have a single overarching seasonality but are influenced by several shifting drivers. Its associated heatmap has an even more scattered pattern, and less possession of defined variations than its best co-cluster counterpart, which means that these are not characteristics enough with discriminatory capabilities to facilitate realistic prediction. Hourly ozone variability plot for its worst co-cluster demonstrates a fall in peak structure, with an even more spiky pattern at unsystematic times around the clock (Fig. 3 (j)). Although there remains some afternoon spike, it is shallow and poorly structured in comparison to the best co-cluster. The broader and less determinate variations suggest that this co-cluster detects cases where ozone generation is influenced by other environmental or meteorological factors, such as wind dispersion or heterogeneous precursor emissions.

The key observation of this comparison is that the best co-cluster utilizes mid-range lag features (lag 13 to lag 17), which well describe ozone trends and permit satisfactory prediction, particularly for the peak summer cases. In contrast, the poorest-performing co-cluster relies on more extended lag periods (lag 18 to lag 24), which fail to establish a strong relationship between historical and present ozone levels, with weaker predictive capability. The disparity in the timing of ozone peaks also works to support the importance of selecting the appropriate window of lags in making forecasts as there are some periods in which the ozone patterns are clearer than others.

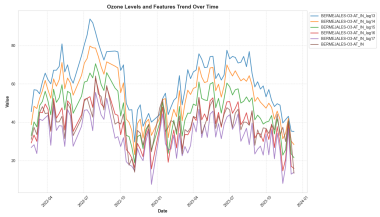
2.4. Explainability Analysis of Best and Worst Co-Cluster of Ronda Del Valle Dataset

The analysis of the Ronda Del Valle dataset's best and worst co-clusters is quite revealing of how different sets of features affect predictions of ozone (O_3) concentrations (Fig. 4). The two clusters include time instances ranging from February 2015 to 2023 and show high ozone levels predominantly in the late afternoon (15:00–20:00) and peak concentrations during the summer (June–August). This trend aligns with atmospheric chemistry principles, as high temperatures and solar radiation enhance photochemical ozone formation. Additionally, increased traffic emissions and industrial activities during these hours contribute to elevated ozone levels. However, the key difference lies in the range of features included in each co-cluster, significantly impacting their predictive power.

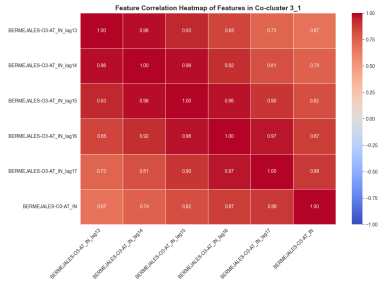
In the best co-cluster, the features include temperature, wind speed, rolling mean, rolling standard deviation, rolling skewness, lag2, lag3, lag4, and lag5 (Fig. 4 (a)). This mixed set of features provides a rich representation of the temporal dependencies of ozone levels and their influencing meteorological factors. The XDeepSpecT heatmap of the best co-cluster indicates structured variability in ozone concentration across different features, which captures meaningful interactions between them. The feature trend graph plots nicely structured periodic variations, particularly for temperature and ozone, that capture the intense diurnal and seasonal fluctuations in the creation of ozone (Fig. 4 (b)). The cyclical patterns in rolling mean, standard deviation, and skewness confirm that these statistical features well represent the behavior of ozone over time, further supporting their utility in predictive modeling. The lag features (lag2 to lag5) are utilized to pick up short-term memory effects, i.e., the model learns about recent ozone variations to provide better predictions. The PCCH of the best co-cluster validates the effectiveness of XDeepSpecT, as shown in the heatmap (Fig.



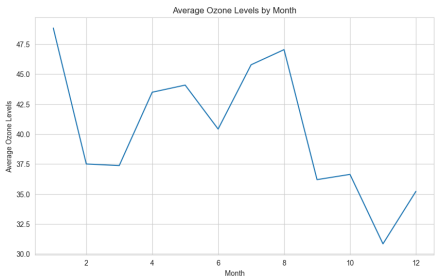
(a) Best co-cluster heatmap



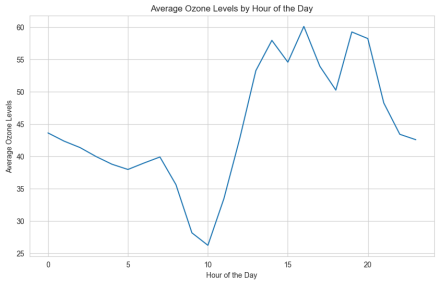
(b) Best co-cluster feature trends



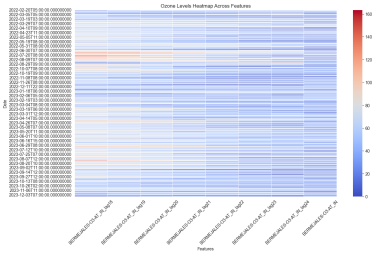
(c) PCCH of features in the best co-cluster



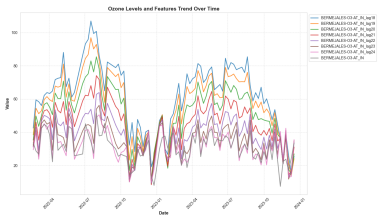
(d) Best co-cluster ozone levels by months



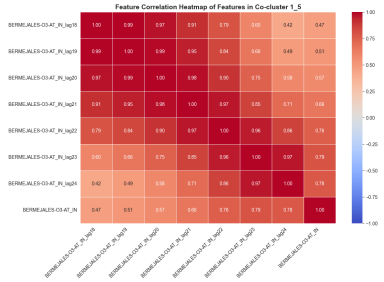
(e) Best co-cluster ozone level by hour of the day



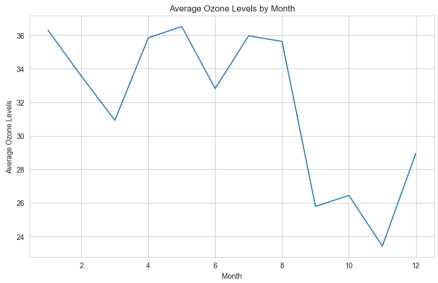
(f) Worst co-cluster heatmap



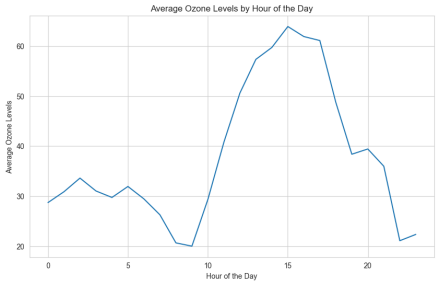
(g) Worst co-cluster feature trends



(h) PCCH of features in the worst co-cluster



(i) Worst co-cluster ozone levels by months



(j) Worst co-cluster ozone level by hour of the day

Figure 3: Analysis of the Best-Performing and Worst-Performing Co-Clusters in the Bermejales Dataset: XDeepSpecT Heatmap, PCCH, Feature Trends, Hourly Variations, and Seasonal Patterns.

4 (a)). It highlights strong correlations among the key features identified by XDeepSpecT, including lag features (lag1 to lag5), rolling mean, and temperature (Fig. 4 (c)).

On the contrary, the worst co-cluster has just one feature: wind speed, which considerably limits its predictive potential. This co-cluster's XDeepSpecT heatmap reveals a much less ordered and more flat pattern of ozone concentrations, reflecting that there are no strong correlations between wind speed in isolation and ozone concentrations (Fig. 4 (f)). The feature trend graph indicates the large variability of the wind speed without any clear periodic structure, making it an unsuitable predictor in isolation for application in forecasting ozone. The absence of other meteorological or historical ozone variables means that the model cannot capture the driving forces of the variability in ozone.

As a whole, the best co-cluster possesses an orderly feature set that excellently extracts ozone behavior by main meteorological and statistical features, yielding cleaner and more stable trends in ozone concentration over time. The worst co-cluster, constrained by a sole feature (wind speed), lacks the information to discern intricate ozone fluctuations, and it shows poor predictive accuracy and more volatile trend patterns. This emphasizes the importance of selecting broad and relevant features for air quality forecasting, particularly in the context of intricate environmental processes such as ozone formation.

2.5. Explainability Analysis of Best and Worst Co-Cluster of Torneo Dataset

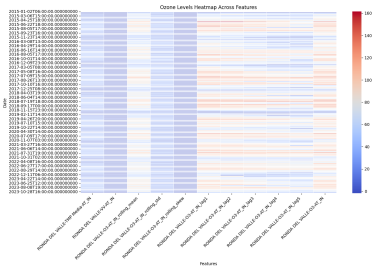
XDeepSpecT approach is used in the Torneo dataset to identify subgroups of data with various temporal patterns and dependencies between the target variable (ozone levels), instances, and features. Fig. 5 summarizes the results for the best and worst co-clusters, emphasizing significant patterns in ozone variations at different temporal scales (daily and monthly) and their dependency on specific features.

In the best co-group (the first two rows) , the XDeepSpecT heatmap highlights the relevance of the selected features in capturing ozone level variations, as evidenced by the pronounced intensity differences between features (Figure 5 (a)). The effectiveness of XDeepSpecT is further validated by the PCCH, which reveals strong correlations between ozone levels and key features—such as temperature, rolling mean, and lagged values (lag19–lag24)—emphasizing their importance in explaining ozone variability (Figure 5 (c)). The temporal trends highlight well-defined diurnal and seasonal cycles, with a sharp rise in ozone levels from May to August (Figure 5(d)), driven by summer conditions that enhance photochemical ozone formation. The diurnal pattern (Figure 5(e)) reveals peak concentrations between 15:00 and 20:00, aligning with the expected late-afternoon ozone increase due to precursor reactions in sunlight.

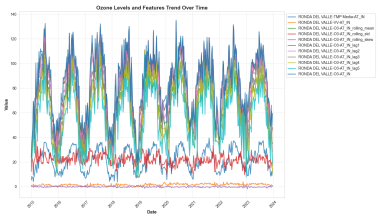
On the other hand, the worst co-cluster (last two rows) exhibits weaker feature-ozone relationships. The heatmap (Figure 5(f)) shows less informative lagged variables (lag7–lag13), resulting in more random fluctuations (Figure 5(g)) and a less structured seasonal rise in ozone from June to August (Figure 5(i)). While the daily cycle (Figure 5(j)) maintains the typical 15:00–20:00 peak, its variability further underscores the poorer predictive power of this co-cluster.

A key difference between these clusters is the feature set used for the prediction of ozone. The most promising co-cluster contains temperature, rolling mean, and lagged terms from lag19 to lag24, allowing it to benefit from both recent and cumulative ozone patterns. Especially important is the inclusion of temperature as a feature, which directly affects photochemical reactions and atmospheric stability and, thus is an important predictor of ozone formation. The rolling mean also contributes predictive power by smoothing out short-term variability and emphasizing longer-term patterns. On the other hand, the worst co-cluster is based on lagged values ranging from lag7 to lag13, which, although offering some history, do not encompass more recent dependencies essential for making accurate ozone predictions. The exclusion of temperature and rolling mean from this cluster detracts from its capacity to model ozone fluctuations adequately.

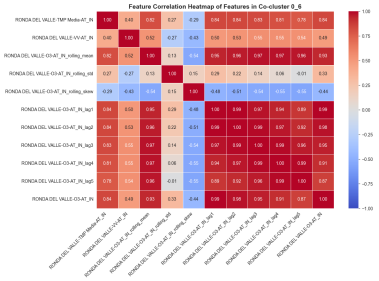
Overall, the best co-cluster demonstrates good predictive capability through the utilization of relevant environmental and temporal features, leading to stronger seasonal and daily trends in ozone levels. The worst co-cluster, constrained by a poorer and less informative set of features, exhibits weaker trends and greater variability in ozone predictions. This difference underscores the importance of selecting appropriate features and meaningful temporal dependencies for modeling the dynamics of air pollution through the application of co-clustering methods.



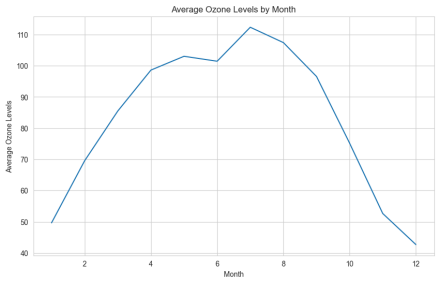
(a) Best co-cluster heatmap



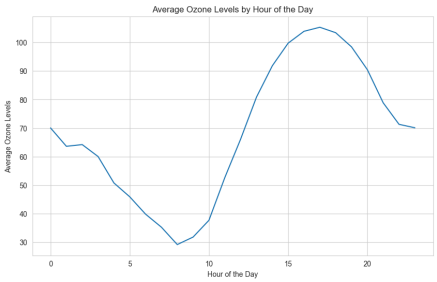
(b) Best co-cluster feature trends



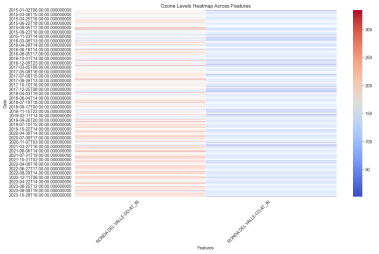
(c) PCCH of features in the best co-cluster



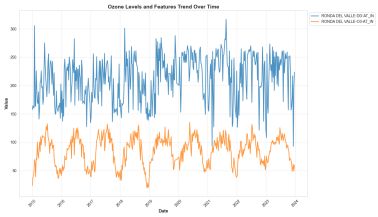
(d) Best co-cluster ozone levels by months



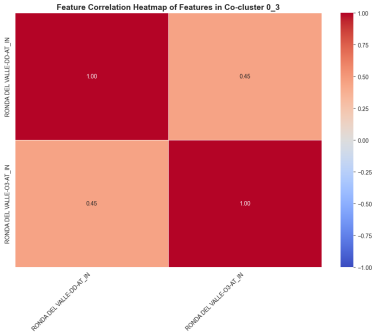
(e) Best co-cluster ozone level by hour of the day



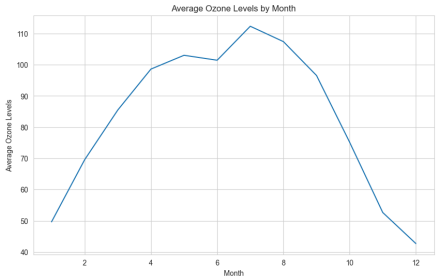
(f) Worst co-cluster heatmap



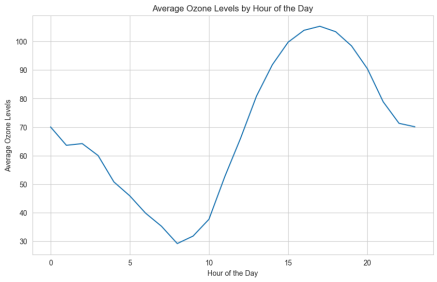
(g) Worst co-cluster feature trends



(h) PCCH of features in the worst co-cluster

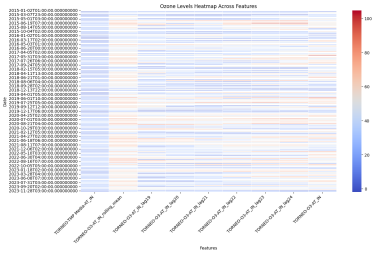


(i) Worst co-cluster ozone levels by months

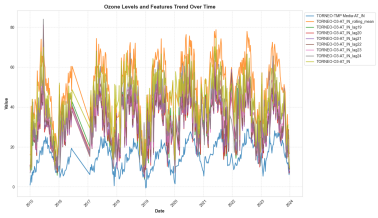


(j) Worst co-cluster ozone level by hour of the day

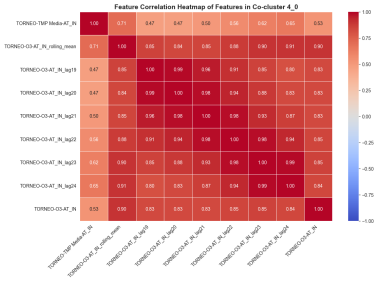
Figure 4: Analysis of the Best-Performing and Worst-Performing Co-Clusters in the Ronda Del Valle Dataset: DeepSpecT Heatmap, PCCH, Feature Trends, Hourly Variations, and Seasonal Patterns.



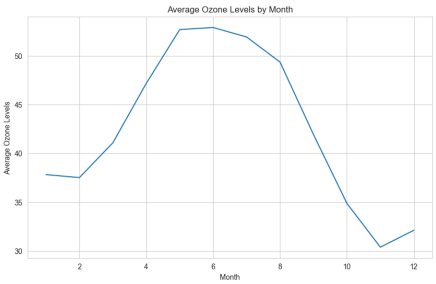
(a) Best co-cluster heatmap



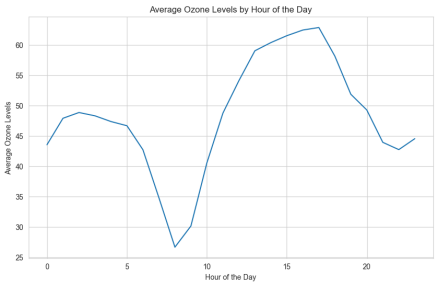
(b) Best co-cluster feature trends



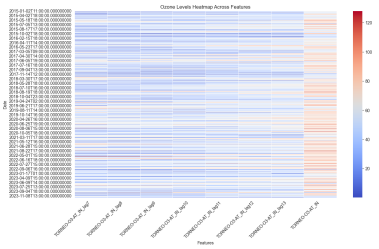
(c) PCCH of features in the best co-cluster



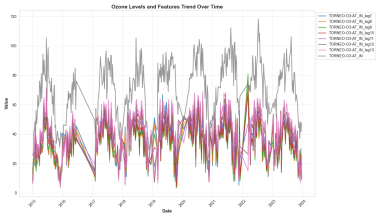
(d) Best co-cluster ozone levels by months



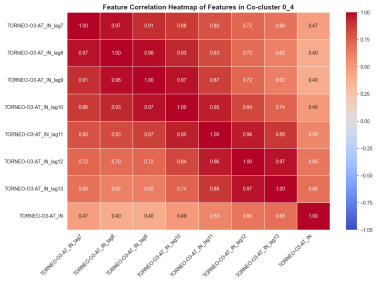
(e) Best co-cluster ozone level by hour of the day



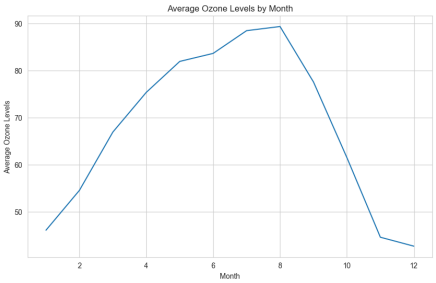
(f) Worst co-cluster heatmap



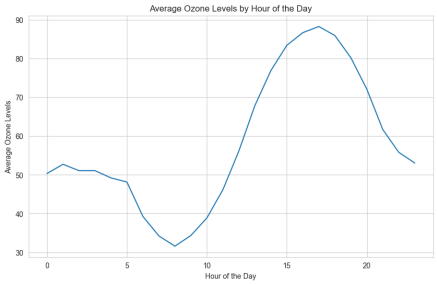
(g) Worst co-cluster feature trends



(h) PCCH of features in the worst co-cluster



(i) Worst co-cluster ozone levels by months



(j) Worst co-cluster ozone level by hour of the day

Figure 5: Analysis of the Best-Performing and Worst-Performing Co-Clusters in the Torneo Dataset: DeepSpecT Heatmap, PCCH, Feature Trends, Hourly Variations, and Seasonal Patterns.