ICA2025OSLO
14th IWA Specialised Conference on
Instrumentation, Control and Automation

# Research on the Performance Improvement of Coagulant Dosing Prediction Models Based on ERA5 Data

Daowei Li*, Mei Han**, Xianyuan Shi, Qinyu Liu, Fang Zou, Ruowei Xie

*lidaowei1993@hotmail.com

** hithanmei@163.com

**Abstract:** Optimal coagulant dosing is critical for water treatment efficiency. Prevailing models relying solely on SCADA data often overlook upstream climate impacts. This study integrates 5-year (2020–2024) SCADA data from a Northern China water supply plant with lagged ERA5 reanalysis fields to predict ferric chloride ($FeCl_3$) and poly aluminum chloride (PAC) dosing using a Bayesian-optimized two-stage XGBoost model. Through 63 experimental scenarios (T0–T62), environmental features—particularly solar radiation, humidity, and evaporation—significantly enhanced performance. The optimal model (T59) reduced mean absolute percentage error (MAPE) by 2.28% (PAC: 7.31%→5.03%) and 3.68% ($FeCl_3$: 10.88%→7.20%) compared to climate-agnostic models. SHAP analysis confirmed PAC' s heightened sensitivity to climate-driven organic load variability, establishing a climate-aware framework for coagulant control.

**Keywords**: ERA5-land; ERA5 reanalysis; Machine Learning; Model Optimization; Coagulant Dosing Prediction Model.

## Introduction

SCADA data (turbidity, pH, temperature, etc.) forms the basis of dosage control but lacks foresight for dynamic conditions (Kyritsakas et al., 2023; Wang et al., 2020). Climate variability profoundly impacts source water quality by altering watershed hydrology, triggering algal blooms, and increasing organic loads (Sharp et al., 2006; Whitehead et al., 2009). ERA5 reanalysis data (Muñoz-Sabater et al., 2021) offers advantages over satellite imagery with its hourly 0.25° resolution, cloud-free continuity, and sub-hourly alignment with SCADA. Despite this potential, integrating ERA5 with SCADA for coagulant prediction remains underexplored. This study bridges this gap by combining operational and climate data from a Northern China water supply plant to quantify environmental contributions to dosing accuracy.

## Materials and Methods

Operational data (2020–2024) came from Plant T's SCADA system (5-minute resolution parameters: flow rate, pH, turbidity, etc.), while ERA5 provided hourly climate variables (humidity, solar radiation, evaporation, etc.) geolocated to Beijing (Hersbach et al., 2020). SCADA data underwent temporal alignment (1-hour resolution), outlier removal, and backfilling. ERA5 variables were processed with 168-hour rolling means to capture lagged environmental effects on algal growth and nutrient runoff (O'Neil et al., 2012). Sixty-three experimental configurations (T0–T62) systematically introduced 0–7 climate variables (Supplementary Table S1). A two-stage XGBoost framework (Han et al., 2022) was implemented (illustrated in Figure 1): Stage 1 predicted $FeCl_3$ dosage using SCADA + ERA5 inputs, while Stage 2 predicted PAC dosage using Stage 1 outputs + SCADA + ERA5. Bayesian hyperparameter tuning (Ankan et al., 2024) optimized learning rate, tree depth, and regularization. Models were evaluated using MAPE, MAE, MSE, $R^2$, and 5-fold time-series cross-validation (Bergmeir & Benítez, 2012), with SHAP analysis (Molnar, 2020) interpreting feature contributions.

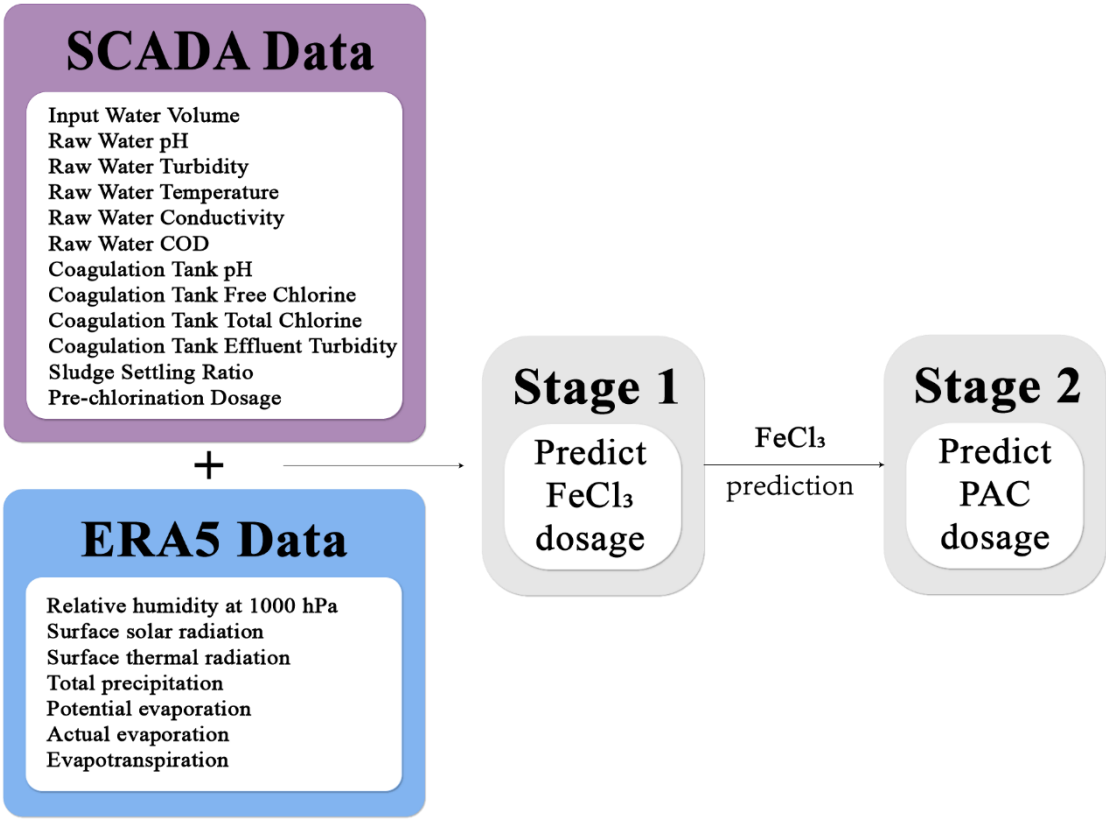


**Figure 1.** Two-stage predictive modeling framework

## Results

Global SHAP analysis (Figure 2) identified raw water turbidity and pH as top features for both coagulants. PAC models showed stronger sensitivity to climate variables (evaporation, solar radiation, humidity), aligning with organic matter dynamics (Stumpf et al., 2020). Figure 3 reveals how feature importance evolved across experimental groups, with climate variables gaining prominence in complex configurations. Optimal performance occurred at T59 (5 variables), reducing PAC MAPE to 5.03% (vs. 7.31% baseline) and $FeCl_3$ MAPE to 7.20% (vs. 10.88%). Performance trends across all configurations (Figure 4) demonstrate significant improvements in $R^2$ (PAC: 0.988→0.994; $FeCl_3$: 0.983→0.992) and diminishing returns beyond five climate variables.

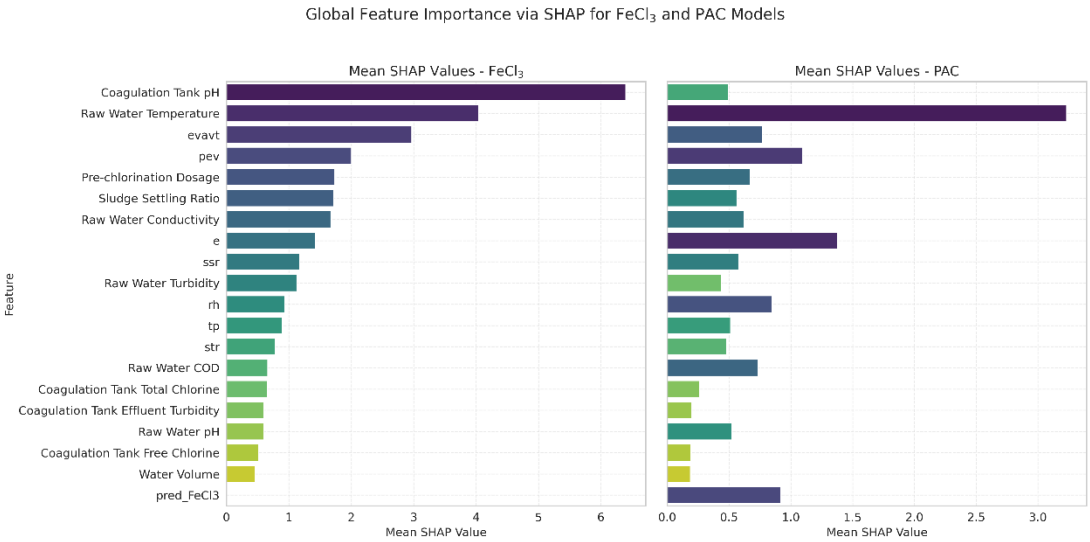

**Figure 2.** Global feature importance via mean |SHAP| values

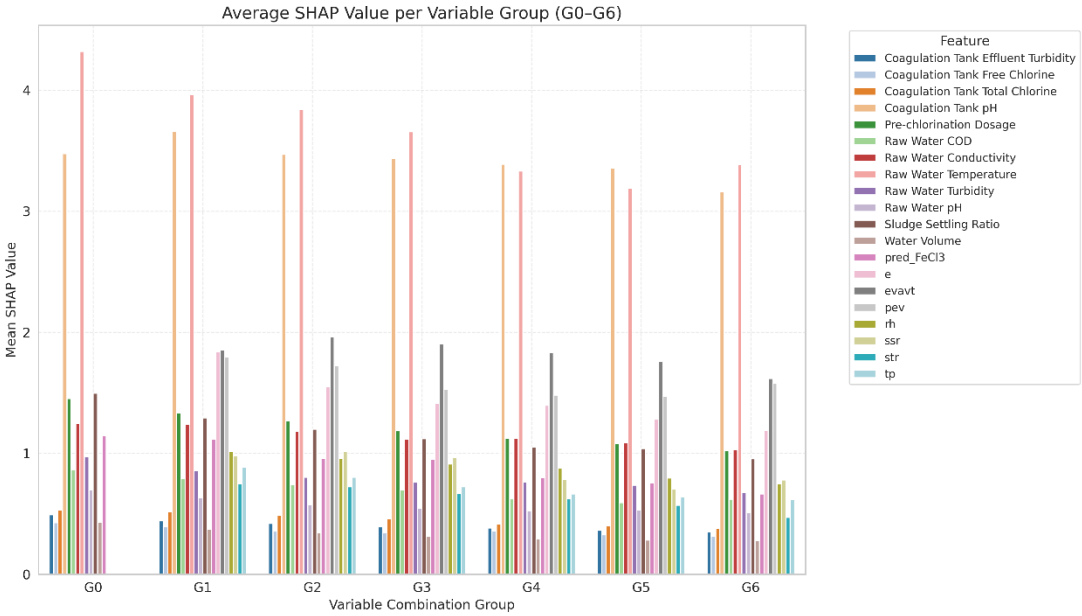


**Figure 3.** Group-wise mean SHAP values

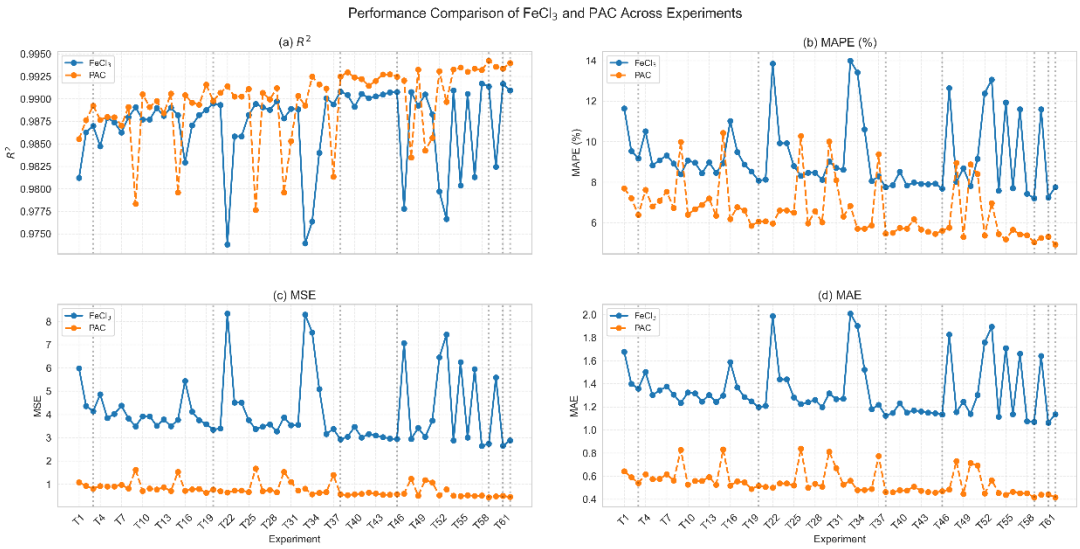


**Figure 4.** Performance metrics across configurations

## Discussion

ERA5 integration captured cumulative environmental impacts (Michalak et al., 2013), with PAC showing greater sensitivity to climate-driven organic variability than $FeCl_3$ (Matilainen et al., 2010). The two-stage architecture (Figure 1) enhanced operational practicality (Han et al., 2022). SHAP visualizations (Figures 2 & 3) confirmed evaporation and solar radiation as key PAC drivers. Performance metrics (Figure 4) validated the optimal 5-variable threshold. Limitations include ERA5's ~31-km resolution (Muñoz-Sabater et al., 2021) and exclusion of extreme events (Medeiros, 2022). Future work should incorporate ERA5-Land, hierarchical SHAP, transfer learning (Newhart et al., 2019), and SCADA cybersecurity protocols (Yadav & Paul, 2020).

**Conclusion**

Lagged ERA5 data integration significantly enhances coagulant dosing predictions in reservoir-sourced water treatment. The Bayesian-optimized two-stage XGBoost model (T59) reduced PAC and $FeCl_3$ MAPE by 2.28% and 3.68%, respectively, identifying solar radiation, humidity, and evaporation as key drivers via SHAP analysis. Performance plateaued at five climate variables, demonstrating a complexity-accuracy trade-off. This framework provides a scalable approach for climate-adaptive water treatment (Bowden et al., 2005), though future efforts should address spatial resolution limitations, extreme event modeling, and operational cybersecurity.

**The significance and impact of your presentation on the topic of the conference**

This research demonstrates significant impact through integrating lagged ERA5 climate data with SCADA operational data, reducing PAC and $FeCl_3$ dosing prediction MAPE by 2.28% (7.31% → 5.03%) and 3.68% (10.88% → 7.20%), respectively, in the optimal configuration (T59). SHAP analysis identified solar radiation, humidity, and evaporation as key environmental drivers of coagulant demand, particularly for PAC sensitivity to organic load variability (Figures 2 & 3). Model performance plateaued at five climate variables (Group G4/G5, Figure 7), revealing a critical complexity-accuracy trade-off. This establishes a climate-aware predictive framework for adaptive water treatment operations, though future work should address spatial granularity limitations (~31 km ERA5 resolution) and incorporate extreme event modeling.

**REFERENCES**

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213.

Han, M., Zou, F., Yuan, R., Xie, R., & Gu, J. (2022). [Discussion on the application effect and implementation path of an intelligent coagulant dosing model based on machine learning]. *Water & Wastewater Engineering, 58*(10), 180–184. Chinese. https://doi.org/10.13789/j.cnki.wwe1964.2022.02.17.0003

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., ... & Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999–2049.

Kyritsakas, K. N., Papafotiou, G. K., Chatzistefanou, A., & Koutsoyiannis, D. (2023). Forecasting bacteriological presence in treated drinking water using machine learning. *Frontiers in Water, 5*, 1199632.

Matilainen, A., Gjessing, E. T., Lahtinen, T., Hed, L., Bhatnagar, A., & Sillanpää, M. (2010). Natural organic matter removal by coagulation during drinking water treatment: A review. *Advances in Colloid and Interface Science, 159*(2), 189–197.

Medeiros, P. M. (2022). The effects of hurricanes and storms on the composition of dissolved organic matter in a southeastern U.S. estuary. *Frontiers in Marine Science, 9*, 841448.

Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N. S., Bridgeman, T. B., ... & Zagorski, M. A. (2013). Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences, 110*(16), 6448–6452.

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.).

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., ... & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data, 13*(9), 4349–4383.

Newhart, K. B., Holloway, R. W., Hering, A. S., & Cath, T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: A review. *Water Research, 157*, 498–513.

O'Neil, J. M., Davis, T. W., Burford, M. A., & Gobler, C. J. (2012). The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae, 14*, 313–334.

Sharp, E. L., Jarvis, P., Parsons, S. A., & Jefferson, B. (2006). Impact of fractional character on the coagulation of NOM. *Colloids and Surfaces A: Physicochemical and Engineering Aspects, 291*(1–3), 237–244.

Stumpf, R. P., Culver, D. A., Tester, P. A., Tomlinson, M., Kirkpatrick, G. J., Pederson, B. A., ... & Litaker, R. W. (2020). Short-term rainfall limits cyanobacterial bloom formation in a shallow lake. *Science of The Total Environment, 700*, 134528.

Wang, D., Shen, J., Zhu, S., & Jiang, G. (2020). Model predictive control for chlorine dosing of drinking water treatment based on support vector machine model. *Desalination and Water Treatment, 173*, 133–141.

Wang, Y., Yang, Q., Liu, Z., Gao, W., & Zhang, J. (2022). Influence of climatic factors on cyanobacteria and green algae distribution in freshwater reservoirs. *Scientific Reports, 12*(1), 1234.

Whitehead, P. G., Wilby, R. L., Battarbee, R. W., Kernan, M., & Wade, A. J. (2009). A review of the potential impacts of climate change on surface water quality. *Hydrological Sciences Journal, 54*(1), 101–123.

Yadav, G., & Paul, K. (2020). Architecture and security of SCADA systems: A review. *arXiv preprint* arXiv:2001.02381.

**Supplementary**

**Table S1. Environmental features used in each experiment (T0–T62).**

| Experiment | ERA5 Features (Full Names) |
|---|---|
| T0 | None |
| T1 | Surface Solar Radiation (ssr) |
| T2 | Surface Thermal Radiation (str) |
| T3 | Actual Evaporation (e) |
| T4 | Total Precipitation (tp) |
| T5 | Potential Evaporation (pev) |
| T6 | Evapotranspiration (evavt) |

| | |
|---|---|
| T7 | Relative Humidity at 1000 hPa (rh) |
| T8 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str) |
| T9 | Surface Solar Radiation (ssr), Actual Evaporation (e) |
| T10 | Surface Solar Radiation (ssr), Total Precipitation (tp) |
| T11 | Surface Solar Radiation (ssr), Potential Evaporation (pev) |
| T12 | Surface Solar Radiation (ssr), Evapotranspiration (evavt) |
| T13 | Surface Solar Radiation (ssr), Relative Humidity at 1000 hPa (rh) |
| T14 | Surface Thermal Radiation (str), Actual Evaporation (e) |
| T15 | Surface Thermal Radiation (str), Total Precipitation (tp) |
| T16 | Surface Thermal Radiation (str), Potential Evaporation (pev) |
| T17 | Surface Thermal Radiation (str), Evapotranspiration (evavt) |
| T18 | Surface Thermal Radiation (str), Relative Humidity at 1000 hPa (rh) |
| T19 | Actual Evaporation (e), Total Precipitation (tp) |
| T20 | Actual Evaporation (e), Potential Evaporation (pev) |
| T21 | Actual Evaporation (e), Evapotranspiration (evavt) |
| T22 | Actual Evaporation (e), Relative Humidity at 1000 hPa (rh) |
| T23 | Total Precipitation (tp), Evapotranspiration (evavt) |
| T24 | Total Precipitation (tp), Evapotranspiration (evavt) |
| T25 | Total Precipitation (tp), Relative Humidity at 1000 hPa (rh) |
| T26 | Potential Evaporation (pev), Evapotranspiration (evavt) |
| T27 | Potential Evaporation (pev), Relative Humidity at 1000 hPa (rh) |
| T28 | Evapotranspiration (evavt), Relative Humidity at 1000 hPa (rh) |
| T29 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e) |
| T30 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Total Precipitation (tp) |
| T31 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Potential Evaporation (pev) |
| T32 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Evapotranspiration (evavt) |
| T33 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Relative Humidity at 1000 hPa (rh) |
| T34 | Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp) |
| T35 | Surface Thermal Radiation (str), Actual Evaporation (e), Potential Evaporation (pev) |
| T36 | Surface Thermal Radiation (str), Actual Evaporation (e), Evapotranspiration (evavt) |
| T37 | Surface Thermal Radiation (str), Actual Evaporation (e), Relative Humidity at 1000 hPa (rh) |
| T38 | Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev) |
| T39 | Actual Evaporation (e), Total Precipitation (tp), Evapotranspiration (evavt) |
| T40 | Actual Evaporation (e), Total Precipitation (tp), Relative Humidity at 1000 hPa (rh) |
| T41 | Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt) |
| T42 | Total Precipitation (tp), Potential Evaporation (pev), Relative Humidity at 1000 hPa (rh) |
| T43 | Potential Evaporation (pev), Evapotranspiration (evavt), Relative Humidity at 1000 hPa (rh) |

| | |
|---|---|
| T44 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp) |
| T45 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Potential Evaporation (pev) |
| T46 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Evapotranspiration (evavt) |
| T47 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Relative Humidity at 1000 hPa (rh) |
| T48 | Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev) |
| T49 | Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Evapotranspiration (evavt) |
| T50 | Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Relative Humidity at 1000 hPa (rh) |
| T51 | Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt) |
| T52 | Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Relative Humidity at 1000 hPa (rh) |
| T53 | Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt), Relative Humidity at 1000 hPa (rh) |
| T54 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev) |
| T55 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Evapotranspiration (evavt) |
| T56 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Relative Humidity at 1000 hPa (rh) |
| T57 | Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt) |
| T58 | Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Relative Humidity at 1000 hPa (rh) |
| T59 | Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt), Relative Humidity at 1000 hPa (rh) |
| T60 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt) |
| T61 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Relative Humidity at 1000 hPa (rh) |
| T62 | Surface Solar Radiation (ssr), Surface Thermal Radiation (str), Actual Evaporation (e), Total Precipitation (tp), Potential Evaporation (pev), Evapotranspiration (evavt), Relative Humidity at 1000 hPa (rh) |