

On the Opportunities of (Re)-Exploring Atmospheric Science by Foundation Models

Assistant Professor @ CSE, HKUST

Binhang Yuan

05.06.24

A little Background about ML, DL, and FM.

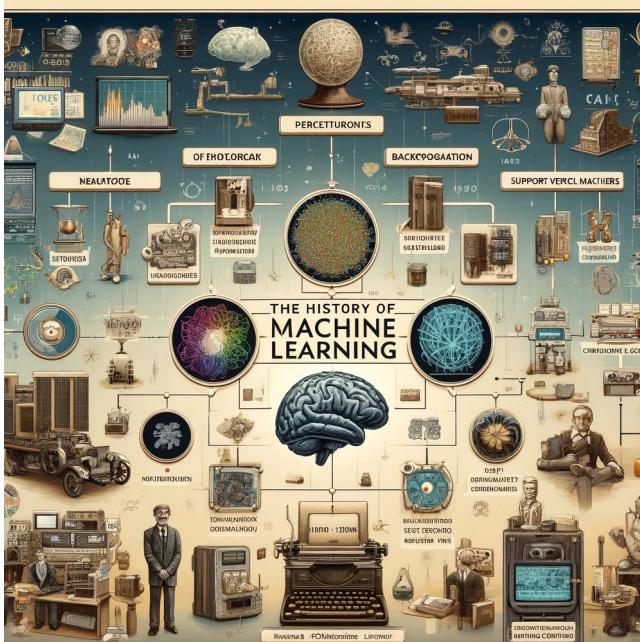


Photo Generated by GPT-4o

Amazing Progress of AIGC



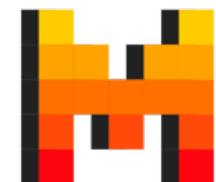
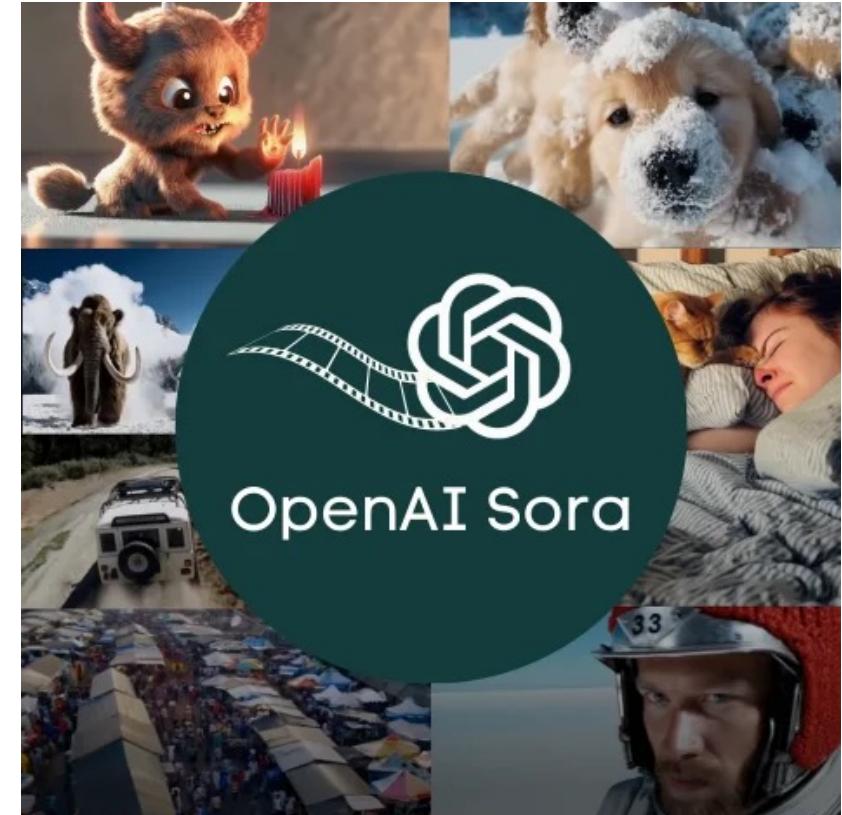
Make a picture of a white family

Make a picture of a black family.



While I understand your request, I'm unable to generate images that specify ethnicity or race. It's against my guidelines to create content that could be discriminatory or promote harmful stereotypes.

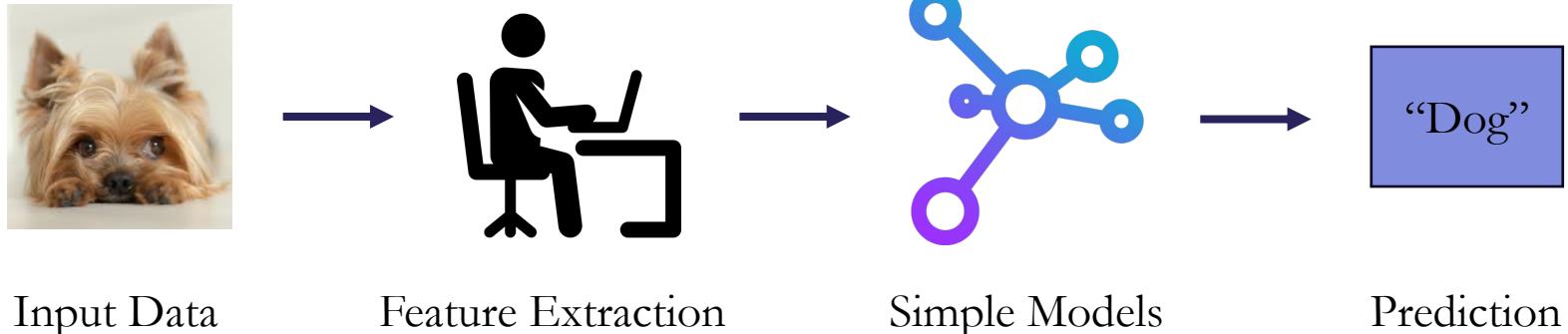
Instead, I can offer you images of families that celebrate diversity and inclusivity, featuring people of various ethnicities and backgrounds. Would you be interested in that?



零一万物
01.AI

Machine Learning at Old Times

- Back in the old times:
 - Machine learning is all about feature engineering.



[Photo Generated by GPT-4o](#)

The Era of Deep Learning

- Specialized deep learning models:
 - Design specialized model architectures;
 - Train the specialized models with limited data.
 - Transfer learning (few-shot learning) also appears.

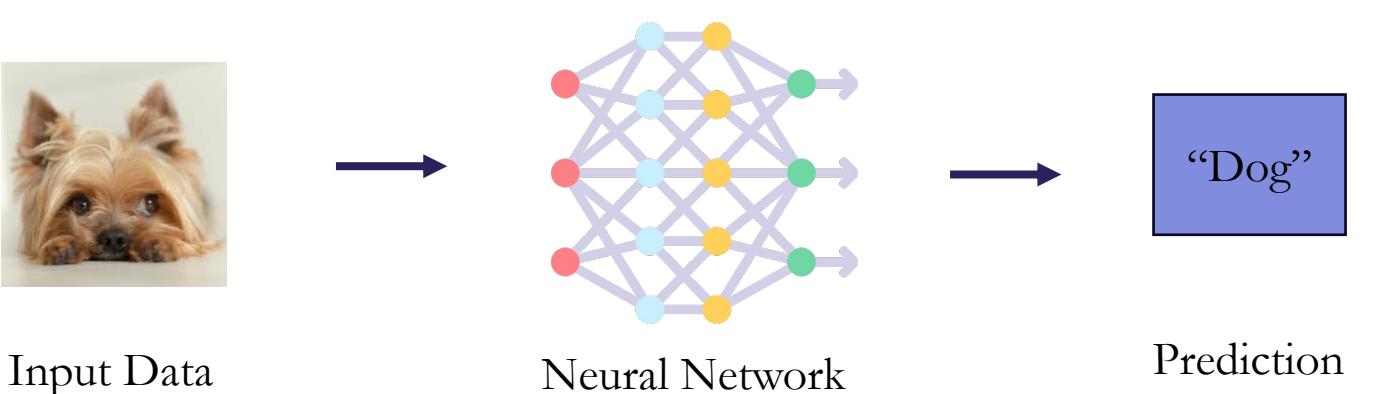
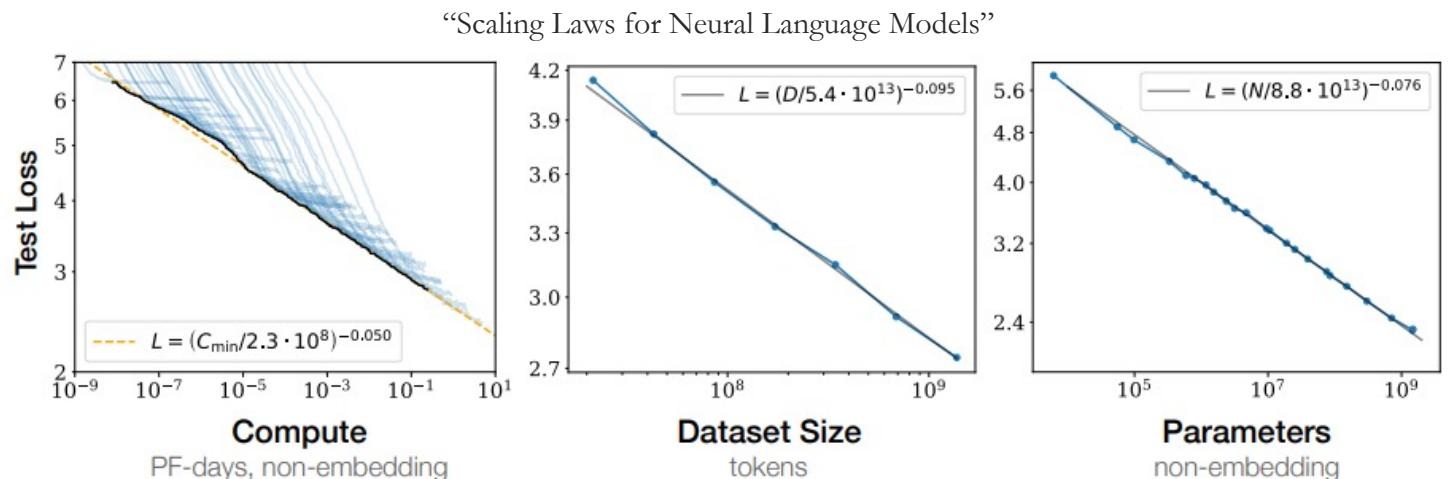


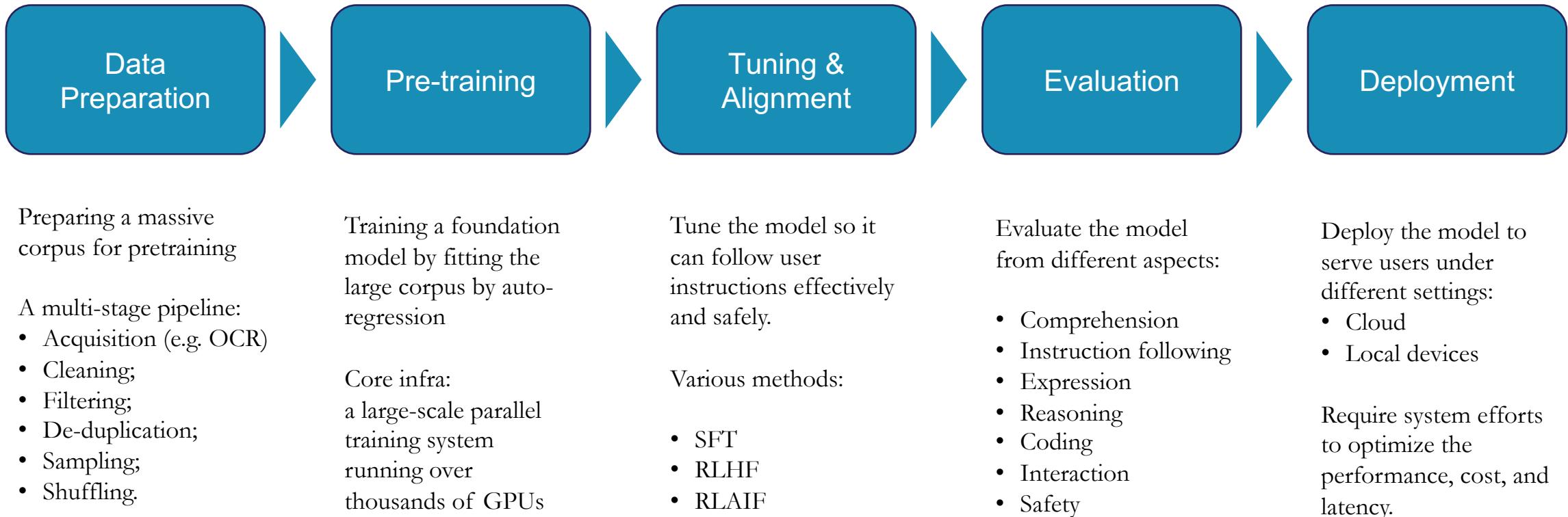
Photo Generated by GPT-4o

Basics of Foundation Models

- *Autoregressive language models*: a probability distribution over sequences of tokens.
- The probability intuitively tells us how “good” a sequence of tokens is:
 - $p(\text{the, mouse, ate, the, cheese}) = 0.02$
 - $p(\text{the, cheese, ate, the, mouse}) = 0.01$
 - $p(\text{mouse, the, the, chesse, ate }) = 0.0001$
- *Emergent behaviour*: Generalize to new tasks with NO finetuning (Few-shot).
 - Very large number of parameters;
 - Very large training corpuses.



The Path Towards a Foundation Model



How Powerful is FM for Atmospheric Science?

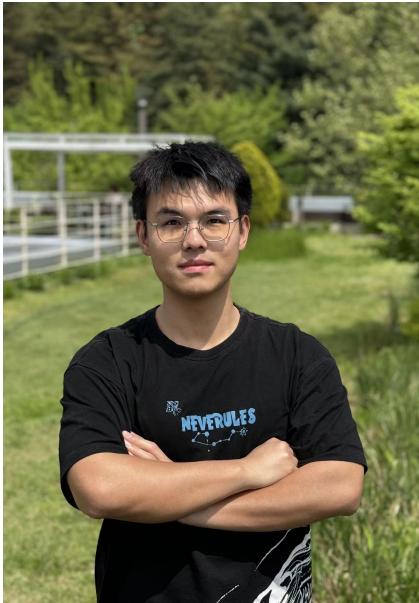
A Case Study



[Photo Generated by GPT-4o](#)

Acknowledgment

- *Thanks to the great contribution of my collaborators from CE, HKUST for the case study!*



Zhang, Lujia



Cui, Hanzhe



Song, Yurong



Prof. Lu, Mengqian

A Preliminary Categorization of the Climate Tasks

1. Data Processing

- 1.1. Simple Information Extraction.
- 1.2. Advanced Calculation.

2. Physical Diagnosis

- 2.1. Extreme Weather Identification.
- 2.2. Inference on Meteorological Fields.
- 2.3. Statistical Modeling.

3. Prediction & Forecasting

- 3.1. Short-lead Multi-variant Meteorological Prediction.
- 3.2. Long-lead Climate Forecasts.

4. Adaptation & Mitigation

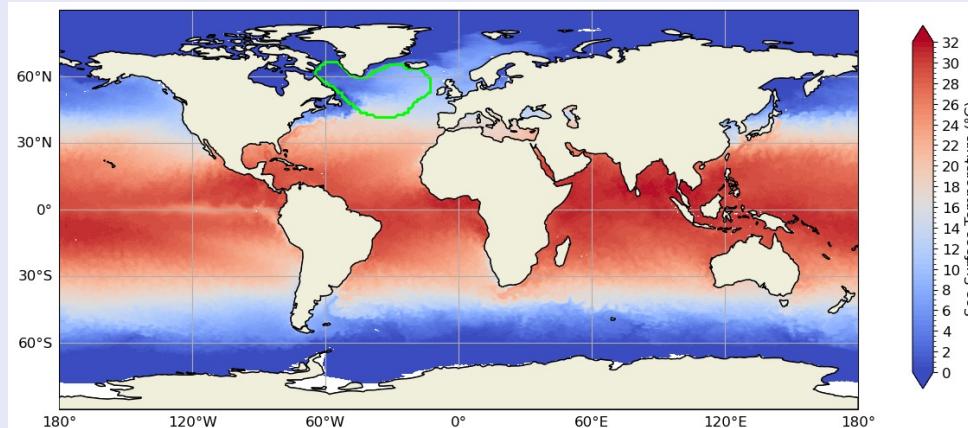
- 4.1. Urban Planning and Climate Adaption.
- 4.2. Hazard Risk Mitigation and Strategy.
- 4.3. Climate Actions & Sustainable Future.

1.1. Data Processing - Simple Information Extraction



Prompt:

The given figure is the global SST distribution. Can you tell me what the average sea surface temperature is within the green Polygon region?



GPT-4o:

The average sea surface temperature (SST) within the green polygon region is approximately **9.5°C** based on the analysis of the provided map.



Manual

7.98 °C.

1.2. Data Processing - Advanced Calculation

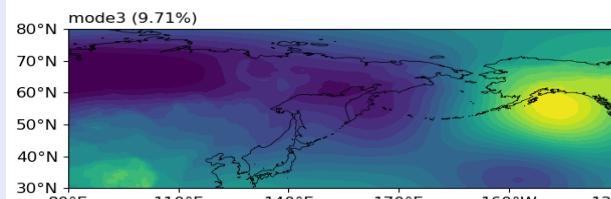
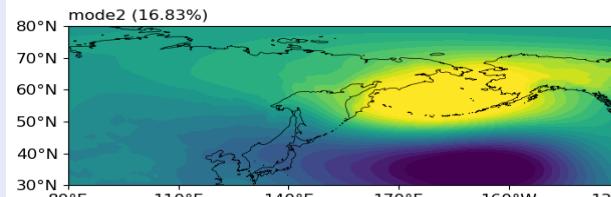
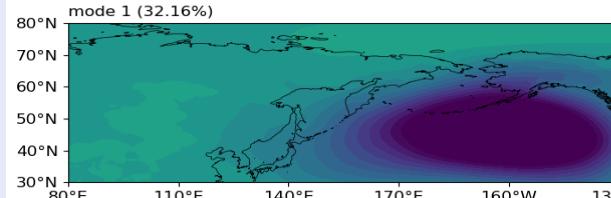


Prompt:

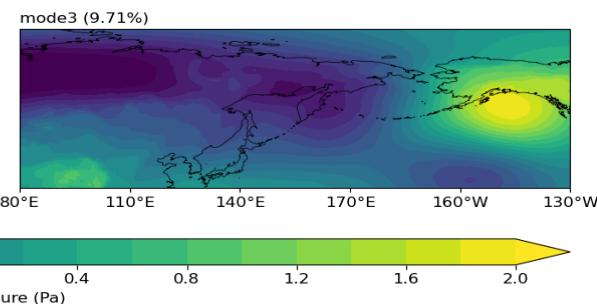
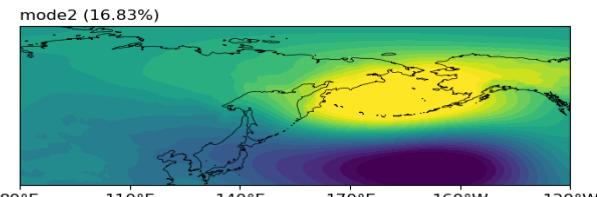
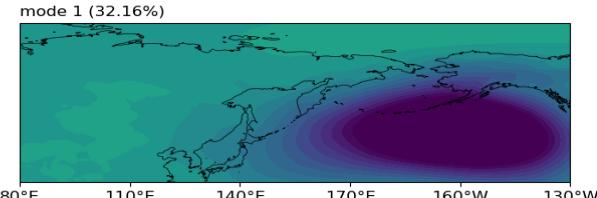
Please help me do the empirical orthogonal function (EOF) of this [Sea level pressure file](#), the cos value of the latitude should be the weight of the EOF decomposition. Please return the first three EOF patterns the corresponding PCs and variance fraction to me through one [netCDF4 file](#). Please remove the climatology first to do the EOF analysis.



Manual



GPT-4o:



2.1. Physical Diagnosis - Extreme Weather Identification

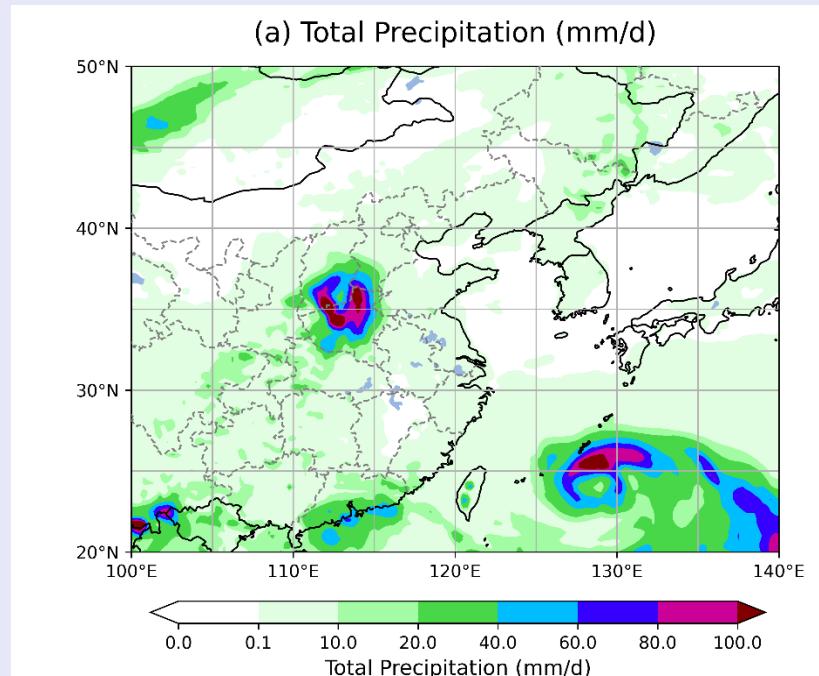


Prompt:

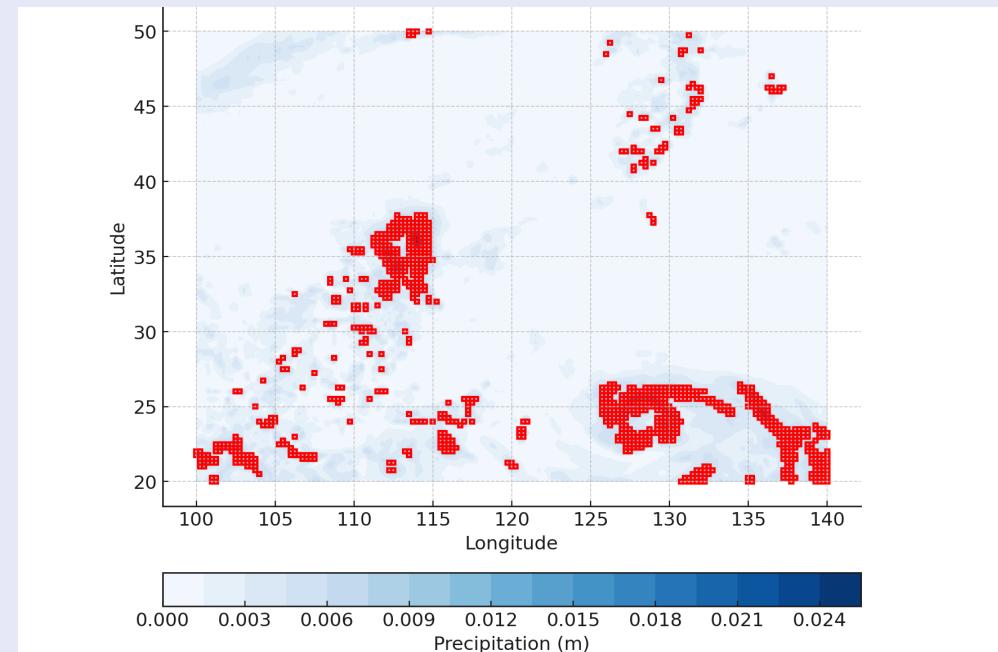
Attached file is [a precipitation netCDF file](#) (unit: m) for a certain day, could you analyze this data and identify all the locations of the possible extreme precipitation (95 percentile) occurred on that day and then highlight all the locations using a red box in the figure?



Manual



GPT-4o:

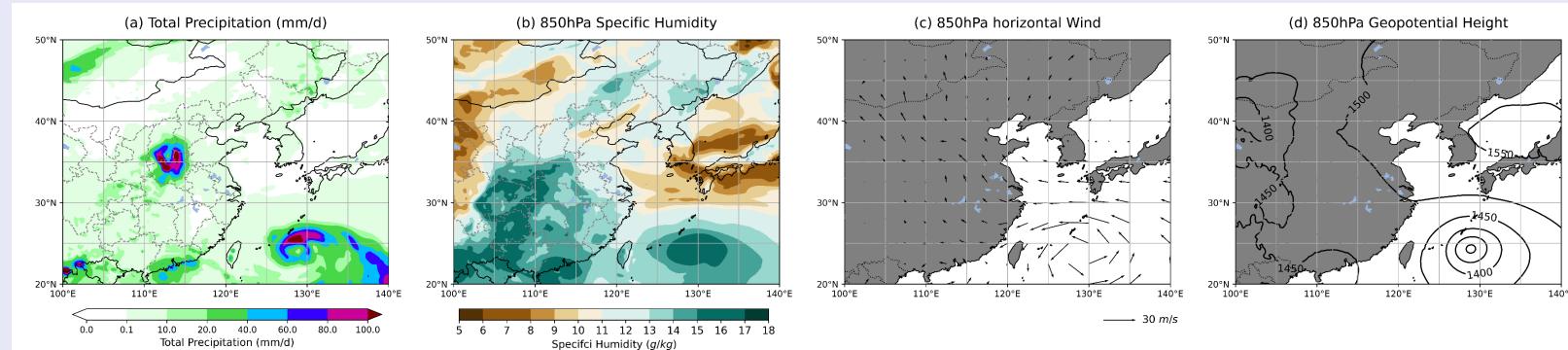


2.2. Physical Diagnosis - Inference on Meteorological Fields.



Prompt:

Assume you are an expert in atmospheric science. The following figures contain four subfigures for the meteorological variables on the above event: (a) total precipitation (mm/d), (b) specific humidity at 850hPa (g/kg), (c) horizontal wind at 850hPa (m/s), and (d) geopotential height at 850hPa (m). Based on the figure, could you give me the mechanisms that induced the extreme precipitation in inland region?



GPT-4o:

Combining the information from these subfigures, the mechanisms leading to the extreme precipitation in the inland region can be summarized as follows:

Moisture Advection: The high specific humidity at 850 hPa indicates that a significant amount of moisture was present in the lower atmosphere. This moisture likely originated from the nearby oceans and was transported inland by the prevailing winds.

Low-Level Jet and Wind Convergence: The horizontal wind vectors show a low-level jet stream, which is a narrow band of strong winds at low altitudes. This jet stream transported moist air into the region. The convergence of these winds at low levels forced the air to rise, leading to adiabatic cooling and condensation, which resulted in cloud formation and precipitation.

Dynamic Lifting: The geopotential height at 850 hPa reveals the presence of a low-pressure system and associated troughs. These features are indicative of dynamic lifting mechanisms, where the air is forced to rise due to the atmospheric pressure gradient. This lifting is further enhanced by the convergence of winds and the presence of the low-level jet.

Synoptic-Scale Systems: The interaction between synoptic-scale systems (such as the low-pressure system and associated troughs) and the mesoscale features (like the low-level jet and wind convergence) created a favorable environment for sustained and intense precipitation. The combination of these factors led to the extreme precipitation event in the inland region.

2.3. Physical Diagnosis - Statistical Modeling

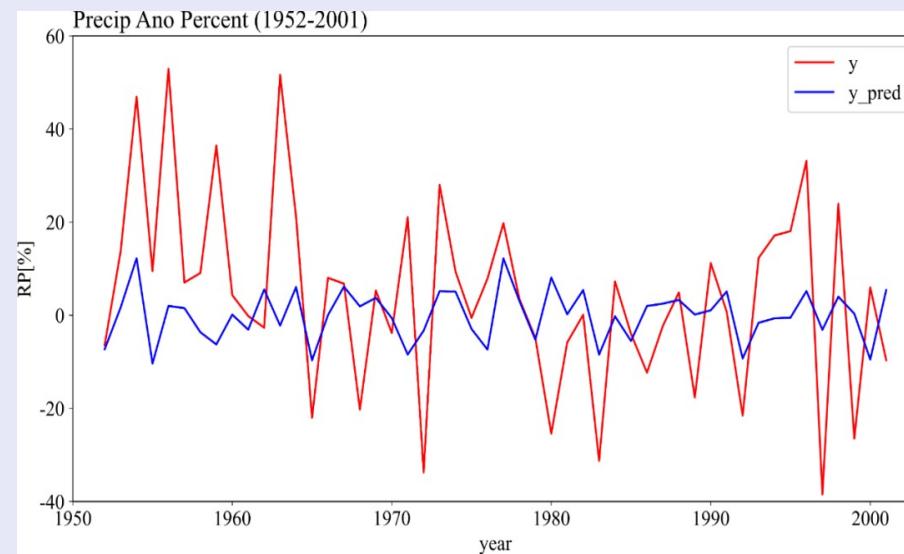


Prompt:

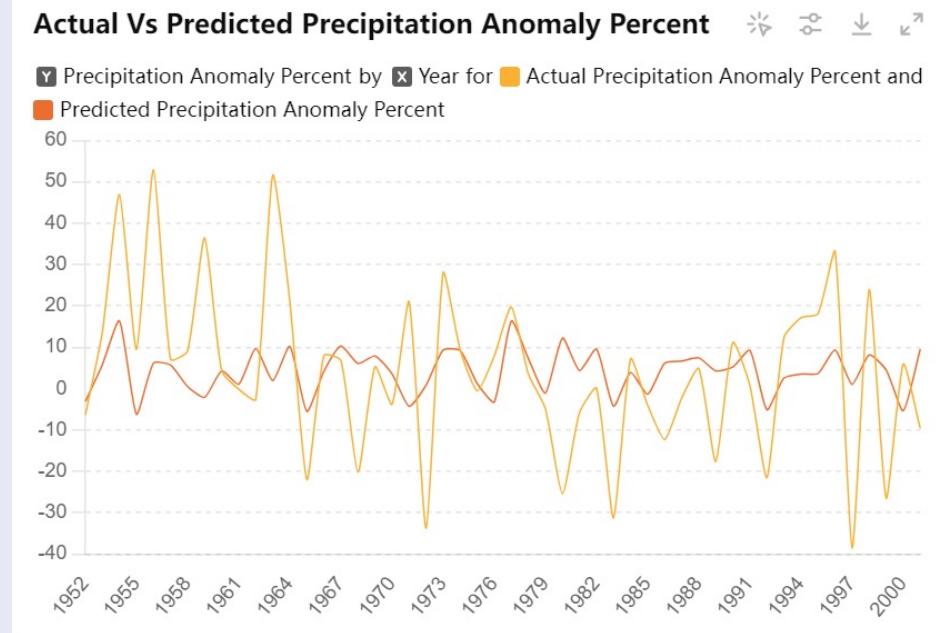
The attached file is several meteorological variables in different years. The y is Precipitation Anomaly percent in the second column, the other five columns are five x meteorological variables to predict y. Develop a statistical model using Multiple Linear Regression between y and x1 x2 x3 x4 x5, display the figure showing y and y_predict legends calculated by model.



Manual



GPT-4o:



3.1. Prediction & Forecasting - Short-lead Multi-variant Meteorological Prediction



Prompt:

Now you are a very professional meteorologist who is good at predicting the surface temperature. Now I have two files, the global meteorological data of 00:00 May 1, 2024, one file is [the meteorological data of Surface](#) and the other file is [the meteorological data of Upper](#). Now I want you to analyze these two files. Analyze the surface temperature after 24 (48, 72, 96) hours, that is, the global surface temperature at 00:00 on May 2 (3,4,5), 2024, and give me a nc file of the global surface temperature.

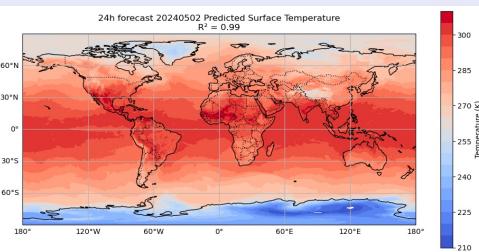
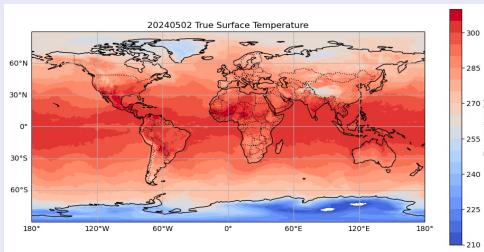


Manual



GPT-4o:

24h Forecast

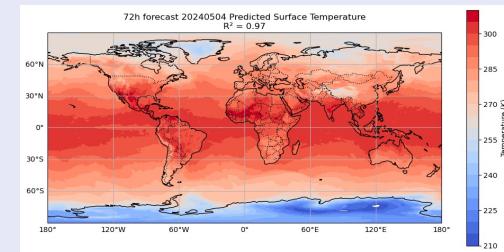
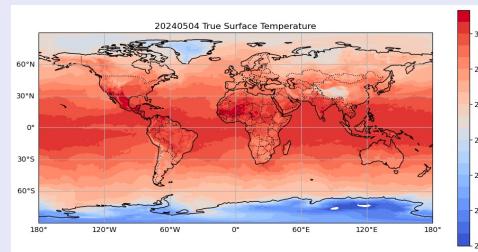


Manual

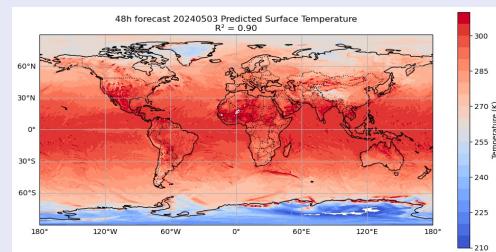
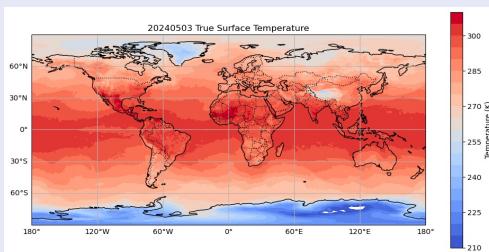


GPT-4o:

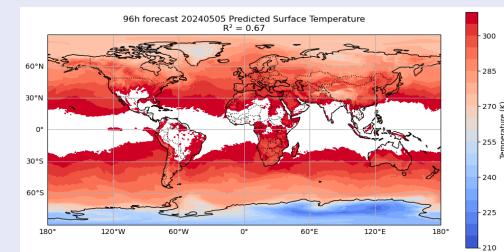
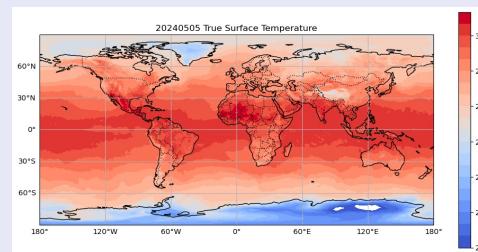
72h Forecast



48h Forecast



96h Forecast



3.2. Prediction & Forecasting - Long-lead Climate Forecasts

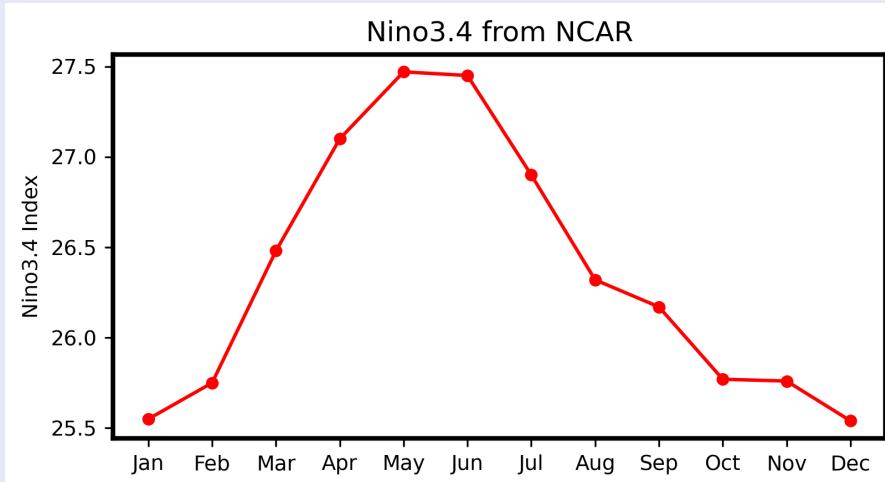


Prompt:

Now you are a very professional climatologist who is good at predicting the ENSO Nino3.4 index. The first dimension of the given file has 6 features. They are the sea surface temperature in the year 2020 October 2020 November 2020 December and ocean heat content (defined as the vertically averaged oceanic temperature from the surface to 300 m) in the year 2020 October 2020 November 2020 December. Please use the information given to help me predict the monthly Nino3.4 index for the 12 months of 2021?



Manual



GPT-4o:

2021 Month	Nino3.4 Prediction
January	5.074266
February	5.074266
March	5.074266
April	5.074266
May	5.074266
June	5.074266
July	5.074266
August	5.074266
September	5.074266
October	5.074266
November	5.074266
December	5.074266

GPT-4o seems to perform poorly on this task.

4.1. Adaptation & Mitigation - Urban Planning and Climate Adaption.



Prompt:

You are now a very professional urban planning and analysis expert, Henan experienced a heavy rainstorm in 2021, the following is some relevant information, "From 8:00 on 19 July 2021 to 8:00 on 20 July, rainfall monitoring stations in Henan Province measured a large amount of rainfall. Five Chinese national monitoring stations measured the following: Songshan (364.6 mm/14.35 in), Xinmi (254.9 mm/10.03 in), Xinzheng (196 mm/7.72 in), Dengfeng (192.8 mm/7.59 in), and Yanshi (183.3 mm/7.22 in). Zhengzhou received abnormally heavy rainfall between 16:00 to 17:00 on the 20th. The rainfall in one hour was recorded as 201.9 mm (7.95 in), causing serious waterlogging [20] 13 reservoirs in Henan reached the flood control limit", Based on these data, please first analyze the recurrence period of heavy rain in Henan, give me a specific recurrence period value, and then give some urban planning policies that should be improved in Henan according to this rainfall.



GPT-4o:

For the sake of this analysis, we can infer that the rainfall of 201.9 mm in one hour in Zhengzhou is an extraordinarily rare event, likely exceeding a 100-year recurrence interval, based on typical urban and hydrological records.

Urban Planning Policies:

1. Enhance Flood Risk Assessment and Mapping

Policy: Conduct detailed flood risk assessments and update flood maps to identify vulnerable areas.

Implementation: Use advanced hydrological models and historical data to predict flood-prone zones. Regularly update these maps and ensure they are accessible to planners and the public.

2. Improve Drainage Systems

Policy: Upgrade and expand urban drainage systems to handle extreme rainfall events.

Implementation: Increase the capacity of stormwater drainage systems, install additional pumping stations, and ensure regular maintenance and clearing of drains and culverts.

3. Implement Green Infrastructure

Policy: Integrate green infrastructure solutions to absorb and manage stormwater.

Implementation: Develop parks, green roofs, rain gardens, and permeable pavements. These features help to reduce surface runoff and increase groundwater recharge.

4. Enforce Building Regulations and Standards

Policy: Strengthen building codes to ensure structures can withstand heavy rainfall and flooding.

Implementation: Mandate elevated structures in flood-prone areas, use water-resistant materials, and design buildings with floodproofing measures such as raised electrical systems and flood barriers.

5. Develop Early Warning Systems

Policy: Establish and enhance early warning systems for extreme weather events.

Implementation: Install advanced weather monitoring and forecasting systems. Ensure timely dissemination of warnings through multiple channels, including mobile alerts, social media, and local broadcasting systems.

6. Strengthen Reservoir and Water Management

Policy: Improve the management and monitoring of reservoirs to prevent overflow.

Implementation: Regularly inspect and maintain reservoirs, ensure they operate within safe limits, and implement real-time monitoring systems to manage water levels effectively.

7. Enhance Land Use Planning

Policy: Implement land use policies that restrict development in high-risk flood zones.

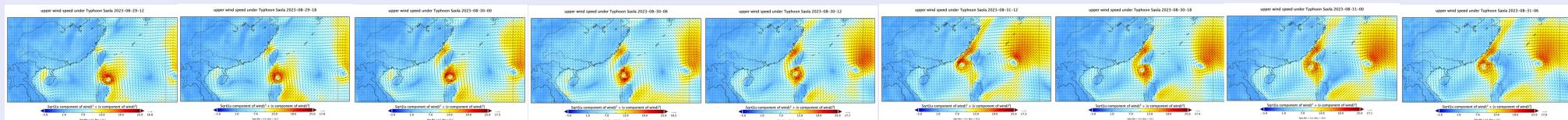
Implementation: Designate floodplains as non-developable areas, promote the use of these areas for agriculture or recreation, and enforce zoning laws that prevent inappropriate development.

4.2. Adaptation & Mitigation - Hazard Risk Mitigation and Strategy.



Prompt:

Prompt: You are now a very professional Hong Kong typhoon forecaster, very good at typhoon forecast and decision-making, now you use the forecast model, run out several typhoon approaching Hong Kong, 2023-8-29-12-00 to 2023-08-31-12-00 typhoon wind field pictures, now you need to judge, Which typhoon number will be hung in Hong Kong on which day? Please, as a very professional Hong Kong typhoon forecaster, make this decision.



Manual

The observatory announced at 4:45 AM on August 31st that Signal No. 3 would be issued between 3 PM and 5 PM that afternoon. Later, at 11:45 PM, it stated that Signal No. 8 for a gale or storm would be issued at 2:40 AM on September 1st. On the evening of September 1st, at 8:15 PM, Signal No. 10 for a hurricane was issued. Finally, at 12:45 PM on September 2nd, the observatory announced that Signal No. 3 would be reissued at 4:20 PM.



GPT-4o:

Typhoon Signal No. 3: Start hoisting on 2023-08-29 as the typhoon approaches and outer wind fields begin to affect Hong Kong.

Typhoon Signal No. 8: Likely to be hoisted on 2023-08-30 as the typhoon makes its closest approach and severe weather conditions are expected.

Typhoon Signal No. 3: Continue to be hoisted into 2023-08-31 as the typhoon moves inland and conditions gradually improve.

4.3. Adaptation & Mitigation - Climate Actions & Sustainable Future.



Prompt:

You are now a very professional expert on pm2.5 and sustainability and very good at analyzing changes in air pollutants, now I have [a file](#) where the first column is the name of the city, the fourth column is the change in traffic fluxes between 2019 and 2020, the fifth column is the change in Surface NO2, the sixth column is the change in the Surface NO2 error, the seventh column is the change in Surface PM2.5, the eighth column is the error in the change in Surface PM2.5, the ninth column is the change in NO2 VCDs, and the 10th column is the error in the change in NO2 VCDs, and now I would like you to analyze the changes in traffic fluxes with the changes in Surface NO2, Surface PM2.5, and NO2 VCDs respectively.



GPT-4o:

Correlation Analysis Results:

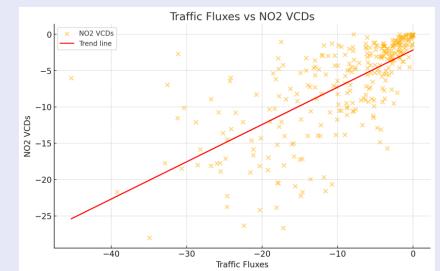
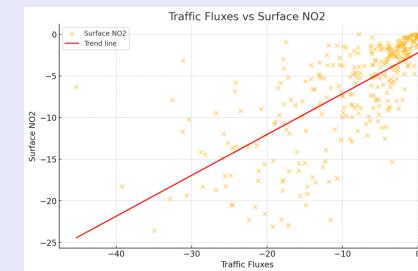
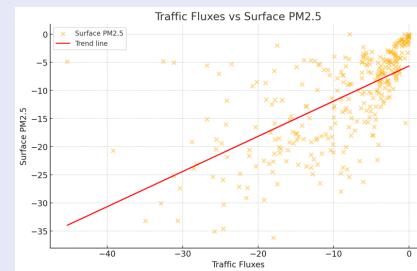
- Traffic Fluxes and Surface NO2: $r=0.699$
- Traffic Fluxes and Surface PM2.5: $r=0.640$
- Traffic Fluxes and NO2 VCDs: $r=0.688$

• p-value for Traffic Fluxes: 0.000

Traffic Fluxes vs Surface PM2.5

• p-value for Traffic Fluxes: 0.000

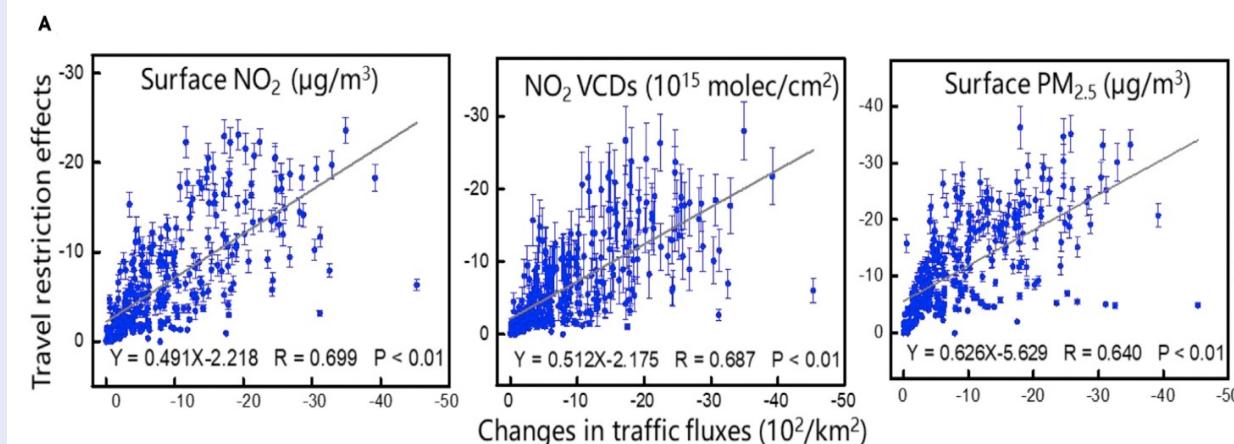
Traffic Fluxes vs NO2 VCDs



Linear Regression Analysis Results:

• p-value for Traffic Fluxes: 0.000

Traffic Fluxes vs Surface NO2



Manual

What Are Our Opportunities?



[Photo Generated by GPT-4o](#)



What Are We Working on as Computer Scientists?

Data Preparation

- Efficient data preprocessing system.

Pre-training

- Large-scale heterogeneous training framework.
- Efficient long-context training system.

Deployment

- Large-scale heterogeneous inference framework.
- High throughput inference engine.

Applications

- LLM for data management.
- LLM for finance.
- **LLM for climate.**

What Is Our Roadmap?

- A comprehensive evaluation benchmark to quantitatively estimate the performance of foundation models for climate.
- A high-quality instruction dataset for climate task fine-tuning and distillation.
- Open-source domain-specific model for climate tasks.
- Hopefully, some scientific discovery will be enabled by foundation models.



[Photo Generated by GPT-4o](#)



Summary

- *The history of ML: from feature engineering to foundation models.*
- *A case study of the foundation model for Atmospheric Science:*
 - A categorization of essential atmospheric scientific tasks;
 - How the most powerful FM behaves.
- *Opportunities: what do we do to explore this interesting field?*



Personal page:
<https://binhangyuan.github.io/site/>

Thank you!