Week 16

Wentao Gao

Research Proposal

- Introduction
- Literature review
- Research questions
- Research methods
- Research plan

Introduction

- Focus the Drought Prediction
- Most prediction models are based on correlation, but the correlation cannot represent the real mechanism of the complex system.
- SO we need causality to help build the prediction model

Literature review

- Dynamic model
- Traditional statistic model
- Machine learning based model
- Deep learning based model
- Causality based model

Dynamic Models

Emphasizing a bottom-up approach, dynamic models focus on simulating the very hydrological processes that underpin droughts.

VIC (Variable Infiltration Capacity) The VIC model primarily simulates the water and energy balance at the surface and near-surface levels. It's especially suited for studying terrestrial hydrological processes.

Noah Land Surface Model: Noah incorporates detailed treatments of soil, vegetation, and snow. It's frequently coupled with climate models and weather forecasting models, providing them with surface boundary conditions

CLM (Community Land Model): CLM is crafted to simulate a spectrum of land processes, including hydrology, ecology, and biogeochemical cycles.

Reference:

1. Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research, 99(D7), 14415-14428.
2. Chen, F., & Dudhia, J. (2001). Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. Monthly Weather Review, 129(4), 569-585.
3. Oleson, K. W., et al. (2010). Technical description of version 4.0 of the Community Land Model (CLM). NCAR Technical Note NCAR/TN-478+ STR.

Traditional statistical methods

Drought predictions heavily relied on time series and regression analyses. These methods utilize historical data, drawing on past patterns and relationships to predict future drought events

Time Series Models:

ARIMA: one of the pioneering efforts, excelled in handling non-stationary data, becoming a mainstay in early drought prediction endeavors (An operational method to forecast reservoir inflow using arima models.)

ETS models(Error, Trends, Seasonality): ETS models further refined this approach, emphasizing the modeling of time depen- dencies. However, their capabilities were often stymied by erratic climatic fluctuations (Forecasting with exponential smoothing: the state space approach.)

The Wavelet Transform: The Wavelet Transform brought a multi-scale perspective, breaking down time series data to unearth patterns at different resolutions. Such an approach proved invaluable for capturing the multifarious nature of hydrological processes (wavelet analysis of river discharge)

Machine learning based method

Feedforward Neural Networks emerged as a simplistic yet adaptable tool, casting a fresh perspective on hydrological patterns, especially in encapsulating concealed, non-linear intricacies prevalent in expansive datasets Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications.

Venturing further into the temporal domain, **Recurrent Neural Networks (RNN)** seamlessly bridged the chasm between static and evolving prediction paradigms, credited to their adeptness at integrating past input sequences(Artificial neural networks as rainfall-runoff models.)

Convolutional Neural Networks (CNNs) was not limited to just image data. They were adapted for sequential precipitation prediction, capitalizing on their ability to detect localized temporal patterns and achieve hierarchical feature extraction in time series data.(An empirical evaluation of generic convolutional and recurrent networks for sequence modeling)

Support Vector Machines (SVMs), with their inherent mathematical elegance, epitomized resilience in data-scarce environments, showcasing an enviable defense against overfitting (Multi-time scale stream flow predictions: The support vector machines approach.)

Deep learning based method

Long Short-Term Memory (LSTM) units stood out as an advanced iteration of RNNs, meticulously catering to long-term sequential data dependencies, while concurrently circumventing the vexing vanishing gradient conundrums that beleaguered their RNN counterparts (Long short-term memory. Neural computation)

Furthermore, the **Transformer** architectures, originally sculpted with linguistic tasks in mind, astonishingly manifested their versatility, adeptly discerning the significance of varied temporal sequences in hydrological data streams Attention is all you need.

Additionally, the Transformer architectures and their derivatives, such as the **Informer, ETSformer, FEDformer** have showcased considerable promise in time series forecasting, optimized to handle hydrological data, thus delivering commendable performances.

Causality based method

- Utilization of techniques like PC and FCI algorithms for uncovering inherent causal structures within large climatic datasets, aiding in understanding the meteorological dynamics related to droughts. More advanced work like PCMCI trying to find the causal relationships among climate features in time series.

- Optimization of feature selection to identify predictors that are statistically significant and causally impactful, providing a stronger foundation for data-driven insights into drought onset and progression.(Causality based feature selection)

- Moreover, within the domain of deep learning, there have been preliminary endeavors to incorporate neural networks with causal pathways, aiming for a more comprehensive understanding of drought dynamics \cite{lee2021}.

Research questions and expected outcomes

- Research Question 1:
- How can causal discovery techniques be leveraged to enhance feature selection, thereby improving the robustness and transferability of drought prediction models.

- Expected Outcome:
- A time-series-oriented causality-driven feature selection algorithm will be developed to identify the salient features that exert genuine influence on the target, hereafter referred to as \ausal features. Represented through a causal graph, these causal features are discerned with heightened predictive capabilities in comparison to existing causal feature selection methodologies.

- Research Question 2: How can a causal feature representation methodology incorporating latent variables be developed to enhance the robustness and generalizability of predictive models?

- Expected Outcome:
- The goal of our research is to design a sophisticated feature representation algorithm that effectively integrates both latent variables and original feature variables to decipher and clarify potential causal relationships. This innovative method aims to tackle the inherent challenges associated with unobserved variables, which might serve as confounders or mediators in drought prediction. By seamlessly merging latent and original feature sets within this algorithm, we strive to comprehensively capture and understand the deep-rooted causal mechanisms influencing droughts..

- Research Question 3: How can causal transfer learning be integrated to optimize domain adaptation and bolster the robustness of a model, ensuring that pre-trained architectures maintain high predictive accuracy on unseen datasets?

- Developing a causality-driven transfer learning methodology that substantially augments the adaptability of predictive models. While its primary application will be in climate science, its design will allow for deployment across various complex systems.

Research Method

For question 1

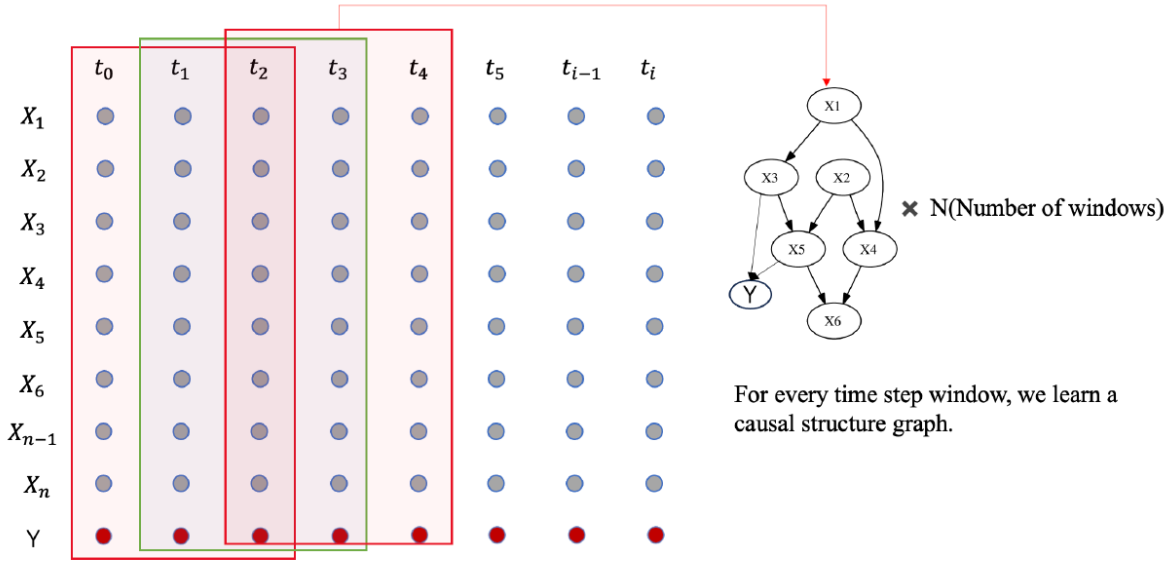

Figure 4.1:

Assumption: causal sufficiency

The DAG (Directed Acyclic Graph) is not merely a visual representation but stands as a practical tool in the feature selection process.

Emphasizing variables that directly influence the onset or severity of droughts allows for a better capture of causality. True causal connections significantly outperform mere correlations in predictive tasks. Once features with high causal relevance are identified, they can be prioritized for model training.

For question 2:

Trying to use ICA as a causal representation learning spirit. There are some works on Nonlinear ICA, Our work is trying to take time delay like x_t-1 into consideration.

For question 3:

Our aim is to amalgamate deep learning with causal modeling, striving to cultivate a deep causal network adept at capturing causal relationships, thereby enhancing both robustness and transferability. Additionally, we envision developing a dynamic domain alignment strategy, capable of real-time adjustments based on fresh data from the target domain, ensuring sustained model performance across evolving landscapes.

Research plan

Year	Time Frame	Phase	Tasks
2023	June - August	Literature Review and Preliminary Planning	- Deep dive into Feature selection,CRL, ICA, and transfer learning - Systematic literature review - Define research questions and objectives
	September - December	Initial Experiments and Data Collection	- Collect and preprocess meteorological data - Preliminary experiments using conventional methods - Initial experiments using causal feature selection
2024	January - April	Model Development - I	- Develop a deep causal network using CRL - Use ICA for feature selection - Start experiments to validate the model
	May - August	Model Development - II	- Implement domain alignment strategy - Continuous validation and model refinement - Start preliminary documentation
	September - December	Deep Evaluation and Refinement	- Comprehensive testing and validation - Identify model strengths and weaknesses - Refine the model
2025	January - June	Incorporating Transfer Learning Techniques	- Integrate causal representation into transfer learning - Develop domain adaptation strategies - Continuous evaluation and refinement
	July - December	Model Optimization and Secondary Testing	- Fine-tune the model - Conduct secondary tests - Start drafting research papers
2026	January - June	Documentation and Paper Writing	- Finalize research findings - Write and refine research papers - Seek expert feedback
	July - October	Final Evaluation and Adjustments	- Conduct final evaluations - Make last-minute model adjustments
	November - December	Submission and Future Work	- Submit research papers - Plan for postdoctoral research