# Causal non-stationarity does not explain seasonal LSTM forecast degradation

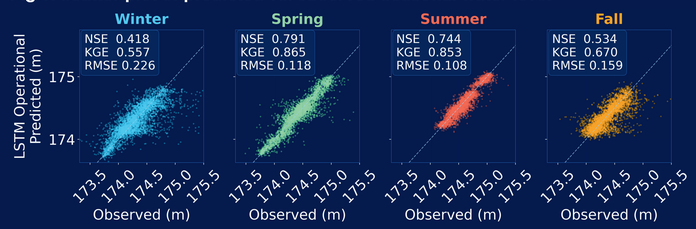
Lucy Myrol[1] and Jan Adamowski[1]

[1]Department of Bioresource Engineering, McGill University, Montréal, H9X 3V9, Canada

## INTRODUCTION

Despite their widespread use in hydrological forecasting, Long Short-Term Memory (LSTM) networks remain challenging to interpret. We find that an LSTM trained for sub-daily water-level prediction in Lake Erie exhibits a pronounced seasonal gradient in skill, for which no simple or widely accepted explanation is evident.

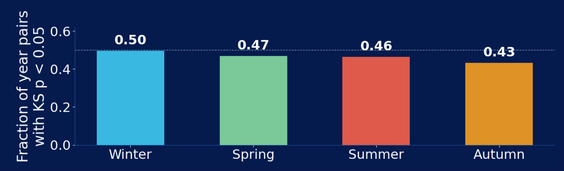
**Fig 1. Scatter plot of predicted vs. observed seasonal water level**




We rule out three common explanations for the observed seasonal performance gradient:

- **Undersampling:** Seasonal representation is similar across the large-sample training dataset (1989 - 2014).
- **Extreme events**: The performance deficit persists across most years and is not driven by extremes.
- **Covariate shift**: Seasonal changes in the marginal input distribution p(x) are no greater in winter or summer than in spring (Fig. 2).

**Fig 2. Kolmogorov-Smirnov test for nonstationarity in the marginal distribution p(x)**




## KEY TAKEAWAY

All seasons are comparably well-sampled, the mean fraction of year pairs exhibiting statistically significant distributional shift in inputs is equivalent across seasons, and the gradient persists across test years.

Ruling out covariate shift motivates an alternative hypothesis in which concept drift, or non-stationarity in the conditional distribution p(y|x), is the governing mechanism for seasonal performance loss. We test this hypothesis using a constraint based causal discovery method to infer the seasonal causal structure governing sub-daily lake level.

## METHODS

PCMCI+ (Tigramite v5.2) is a constraint-based causal discovery algorithm for multivariate, autocorrelated, nonlinear time series.

PC1 Condition Selection → MCI Condititonal Independence Test → Seasonal Causal Graph → Multi-year Aggregation

### Aggregation Strategy

**Harmonic Mean** Combined-evidence consensus using the harmonic mean p-value, with Benjamini–Hochberg (BH) false discovery rate (FDR) correction at q<0.1 across all directed pairs.

### Metrics for quantifying change in the conditional distribution p(y|x)

**CVe** Coefficient of Variation of Effect Sizes

**nSHD** Normalized Structural Hamming Distance

## RESULTS

**Fig 3. Causal graph representing all variables**

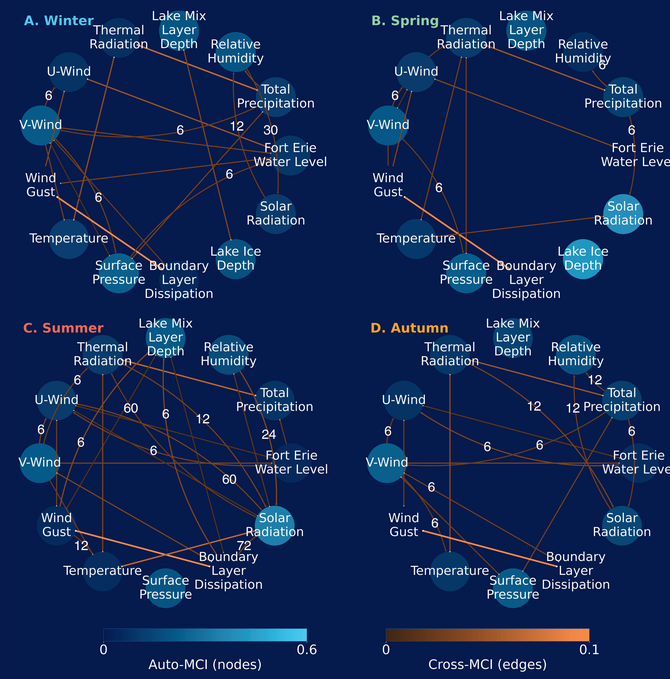



**Fig 4. Causal graph for the parents of lake level only**

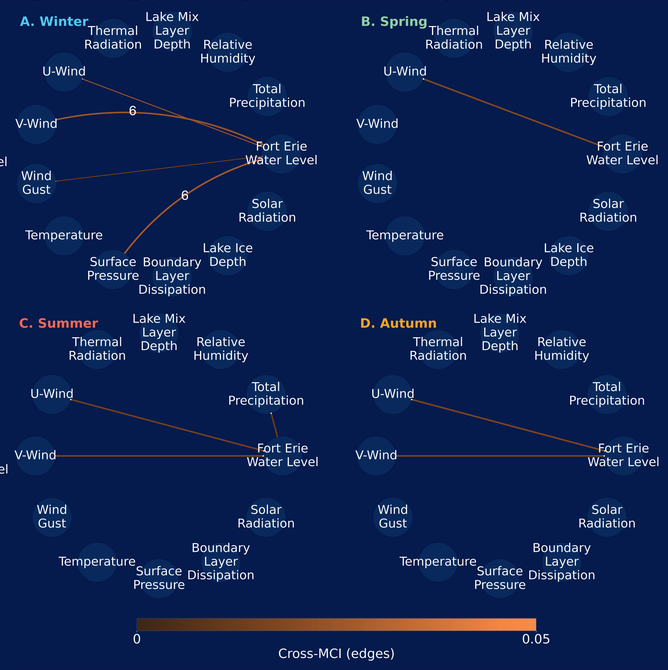



**Fig 5. Seasonal LSTM performance overlaid with interannual variability in nSHD**

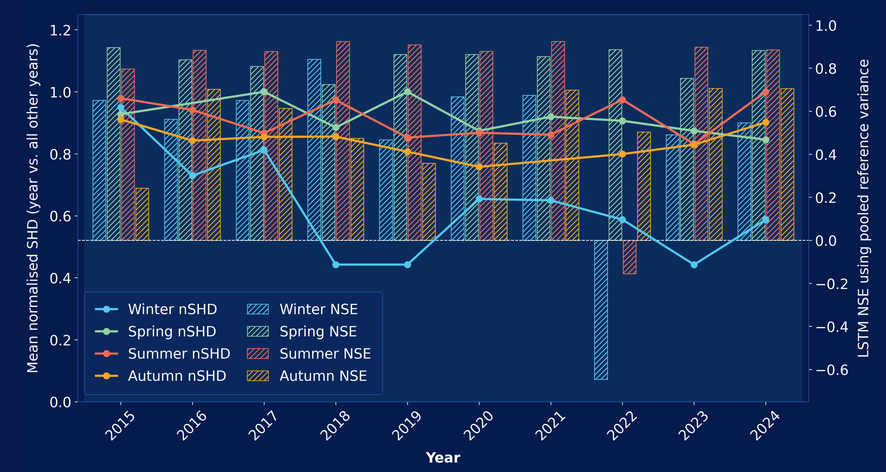



**Fig 6. Count per causal parent of lake level**

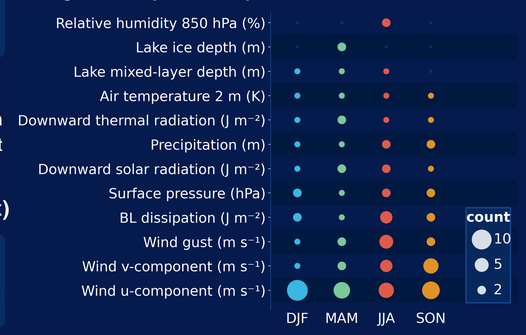
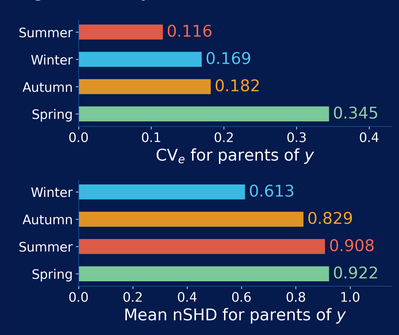



**Fig 7. Instability metrics**




PCMCI+ recovers the causal parents of lake level within each season-year independently. Each seasonal causal graph contains the set of variables that exert a statistically detectable conditional influence on lake level at specific lags.

If the conditional distribution p(y|x) is stationary across years within a season, the same parents should recur with consistent effect sizes. If p(y|x) is non-stationary (i.e. concept drift occurs) the causal graph changes between years. We quantify this change across two levels. Graph instability is assessed using the normalized structural Hamming distance (nSHD), which captures year-to-year changes in the set of inferred causal parents. Mechanism instability is evaluated via the coefficient of variation of effect sizes ($CV_e$), reflecting interannual variability in the strength of persistent causal relationships. Together these metrics function as a proxy for concept drift by operationalizing non-stationarity in the causal graph structure.

## KEY TAKEAWAY

- Graph instability (nSHD) and LSTM performance (NSE) exhibit reversed seasonal rankings.
- Mechanism instability ($CV_e$) is highest in spring, and is otherwise comparable across seasons.

## DISCUSSION

Seasonal differences in LSTM forecast skill were not explained by non-stationarity in the inferred causal structure. Spring showed the greatest concept drift and the highest median forecast skill, while winter showed the most temporally stable causal structure and the lowest forecast skill. Mechanism instability, as measured by $CV_e$, was also highest in Spring.

| Season | NSE Rank | $CV_e$ | nSHD |
|---|---|---|---|
| Spring | 1 | 0.345 | 0.922 |
| Summer | 2 | 0.116 | 0.908 |
| Autumn | 3 | 0.182 | 0.829 |
| Winter | 4 | 0.169 | 0.613 |

The finding that greater instability is associated with higher forecast skill is a counterintuitive result, and suggests that some degree of non-stationarity in the predictor–response relationship is either tolerable to, or potentially exploited by, the LSTM. Generalisation to other hydroclimatic systems remains an open question and requires further validation.

## CONCLUSION

- Evidence of concept drift, defined here as changes in the conditional distribution p(y|x), was strongest in spring and weakest in winter.
- Graph instability (nSHD) and mechanism instability ($CV_e$) were positively associated with LSTM forecast skill.
- Variability in the inferred causal graph is not necessarily detrimental to forecast performance and in some cases may be beneficial.
- PCMCI+-based causal discovery provides a model-agnostic framework for evaluating proxies of concept drift through temporal instability in inferred causal structure.

## REFERENCES

1. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803
2. Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research, 55(12), 11344–11354. https://doi.org/10.1029/2019WR026065
3. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11), eaau4996. https://doi.org/10.1126/sciadv.aau4996

## ACKNOWLEDGEMENTS

We gratefully acknowledge Ontario Power Generation for providing data.