Literature Review

Capstone Projects

Statistical & ML Techniques in Climate Modelling

Karnav Popat

12th Jul ‘24

Paper #1

Review of trend detection methods and their application to detect temperature changes in India (2012)

P. Sonali, D. Nagesh Kumar

Key Methods:

- Parametric slope-based tests (LR & Sen's slope estimator)
- Non-parametric rank-based tests (MK & SRC)
- Adjusting for serial correlation: pre-whitening, variance correction, etc.

Key Findings:

- Assumptions of distribution & independent observations often cause incorrect results.
- Different statistical tests yield different results.
- Serial correlation must be adjusted for while testing and resampling weather data.

Method: Tests

LR test: Standard parametric LR test where slope of the line indicates trend

Mann-Kendall test: Non-parametric test that checks the number of pairs of points where the second point is higher than the first is significantly more than the number of pairs where the reverse is true (or vice versa).

Spearman Rank Correlation test: Non-parametric test that converts observations and timestamps to ranks, then uses the sum of squared differences to calculate a SRCC from [-1, 1] indicating positivity and strength of trend.

Sen's Slope Estimator: Parametric test that takes each possible pair of points and calculates the median of their slopes (the average trend).

Method: Serial correlation effect

TFPW with MK test: De-trend the timeseries and find the lag-1 correlation coefficient of the resulting series. If significant, apply MK test to the de-trended series.

Serial-correlation variance correction: Multiply the variance of the test statistic by a correction factor:

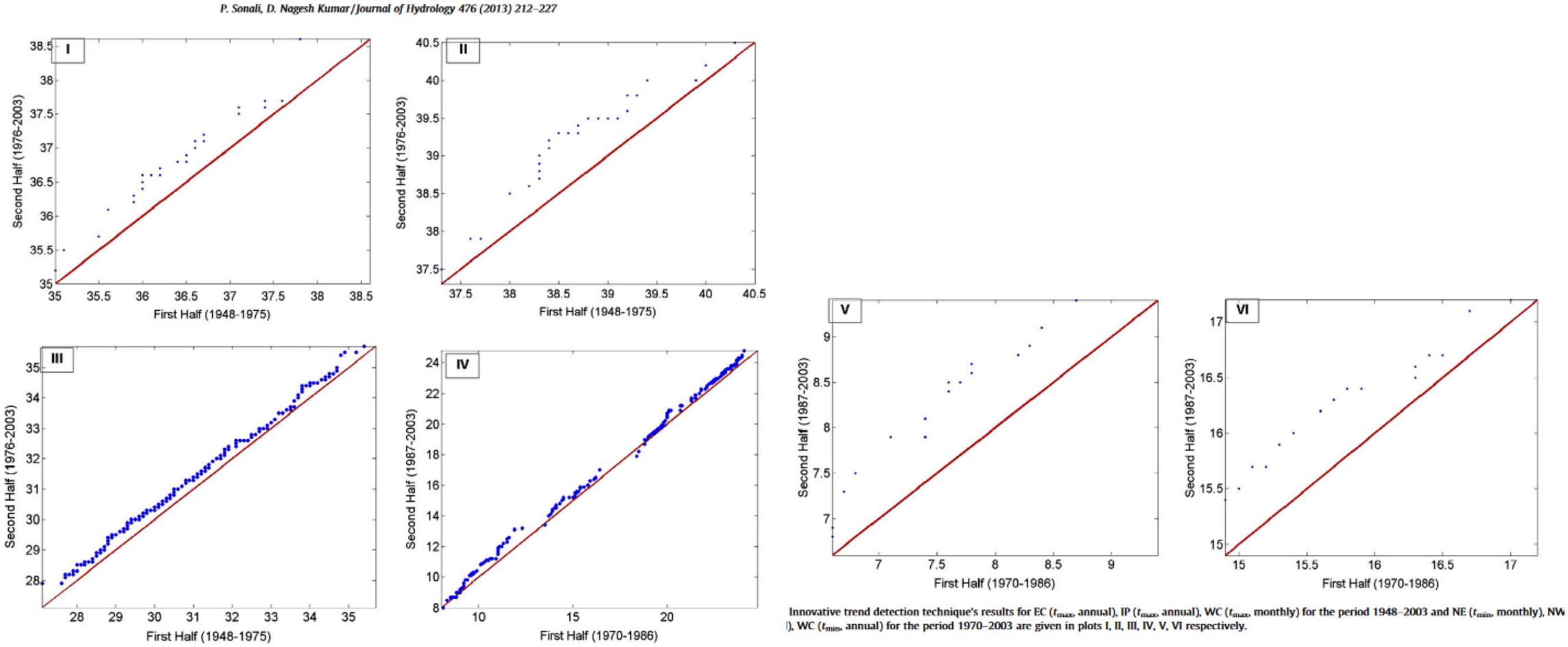
$$CF_1 = 1 + \frac{2}{n(n-1)(n-2)} \sum_{k=1}^{n-1} (n-k)(n-k-1)(n-k-2) r_k^R$$

Block bootstrapping

$$CF_2 = 1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) r_k$$

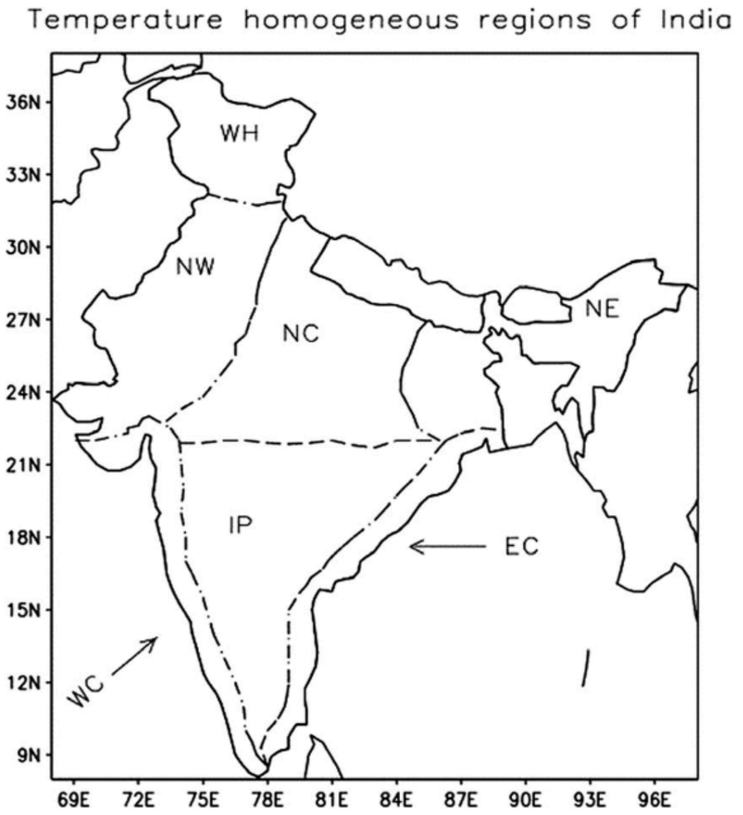
where r_k and r_k^R are the lag-k serial correlation coefficients of data and ranks of data respectively and n is the total length of the series.

Method: Novel trend detection



Innovative trend detection technique's results for EC (t_{max}, annual), IP (t_{max}, annual), WC (t_{max}, monthly) for the period 1948–2003 and NE (t_{min}, monthly), NW), WC (t_{min}, annual) for the period 1970–2003 are given in plots I, II, III, IV, V, VI respectively.

Data & Variables



t_max and t_min IMD-like data divided into three time slots, 1901-2003, 1948-2003 & 1970-2003.

Divides the year into:

- Winter (JF)
- Pre-Monsoon (MAM)
- Monsoon (JJAS)
- Post-Monsoon (OND)

Paper #2

Improved Surface Temperature Prediction for the Coming Decade from a Global Climate Model (2007)

Doug M. Smith, Stephen Cusack, Andrew W. Colman, Chris K. Folland, Glen R. Harris, James M. Murphy

Key Methods:

- Tested Decadal Climate Prediction System for each season for 20 years.
- Measured the impact of initial conditions on prediction

Key Findings:

- Short-term temp changes will be offset by internal variability but will rise long-term.
- Accurate initial conditions have a strong positive impact on predictions.
- Counterintuitively predictions don’t get worse over longer horizons

Method & Model

Uses the Decadal Climate Prediction model built on the Hadley Centre Coupled Atmosphere-Ocean Model 3:

- Grid point model at a resolution of 3.75 × 2.5 degrees (atmosphere) and 1.25 x 1.25 degrees (ocean), with 30-minute timesteps.
- Atmo model is run for one day to generate fluxes (heat, moisture, momentum), then ocean model is run to generate reverse fluxes.

Ran two ensembles initialized on four consecutive days each. DePreSys started from the accurate atmo-ocean conditions on that day, while NoAssim started from an independent set of conditions within the range of variability for that day.

Used RMSE to measure accuracy of predictions for the season based on the ensemble results of the two models.

Data & Findings

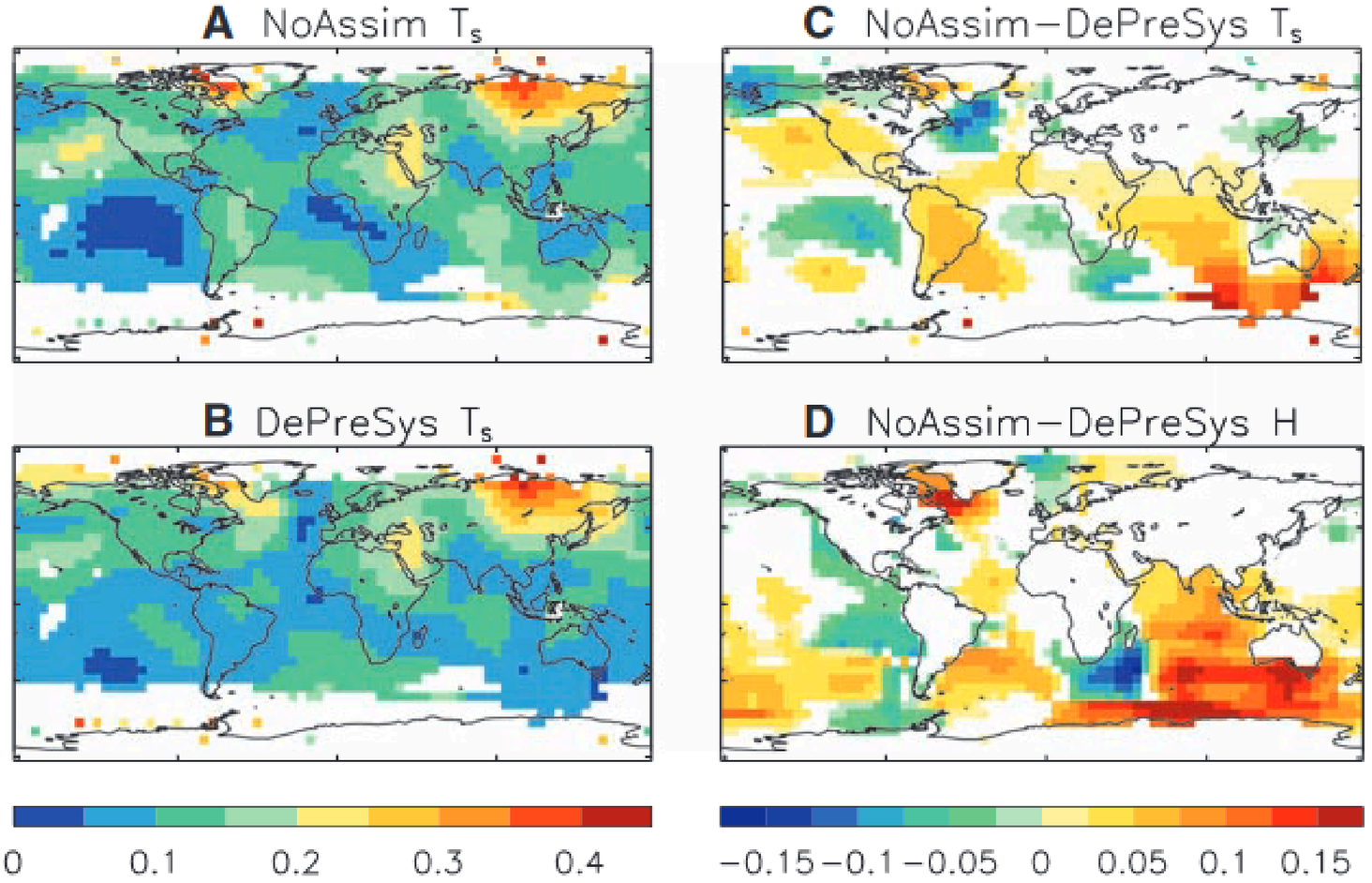
Used the Decadal Climate Prediction model, which has granular data from 1971-2005 for atmospheric temperature, pressure, moisture, and wind velocity.

Split training into 80 start points, one for each quarter between 1982-2001 and calculated the RMSE of predictions for each ensemble.

Accounted for adjustments due to black swan events and unpredictable trends such as El Nino and volcanic eruptions by subtracting differences from the data.

Found that perfect knowledge of initial conditions and more accurate simulation of atmosphere-ocean interactions significantly reduce bias in prediction. This remains true at the near-term annual and long-term decadal scale, for all regions. Lacking initial information, NoAssim failed the worst in open ocean regions.

Findings


A NoAssim T_s
C NoAssim−DePreSys T_s
B DePreSys T_s
D NoAssim−DePreSys H
0 0.1 0.2 0.3 0.4
−0.15 −0.1 −0.05 0 0.05 0.1 0.15

Paper #3

Artificial neural networks for automated year-round temperature prediction (2009)

Brian A. Smith, Gerrit Hoogenboom, Ronald W. McClendon

Key Methods:

- For each NN architecture, 30 models were trained and selected based on MAE
- Time and seasonality were represented by triangular fuzzy logic
- One hidden layer with many activation functions

Key Findings:

- Rainfall is a key indicator
- Cloud cover estimates should be used as inputs
- ML-based methods are more accurate than statistical methods

Method & Model

Multiple NN architectures were shortlisted as candidates. For each architecture:

- 30 feed-forward backprop models were trained with different random initial weights for a 4h horizon.
- Training data was a set of 300,000 patterns for 15 epochs chosen from a development set of 1.2m (1997-2000).
- Each model had one hidden layer with 120 nodes in three equal slabs (hyperbolic tangent, exp(-n^2), and 1-exp(-n^2)) and one output sigmoid node.
- The best model was chosen based on minimum MAE on the development set.

Once the best performing model of each architecture was chosen, it was then tested against the selection set of 1.2m patterns (2001-2003) and the model with the minimum MAE over the selection set was chosen.

Data & Variables

The dataset was drawn from weather data from 21 rural weather stations in Georgia, US (Georgia University AEMN) for 1997-2005, representing 3.3 million data points.

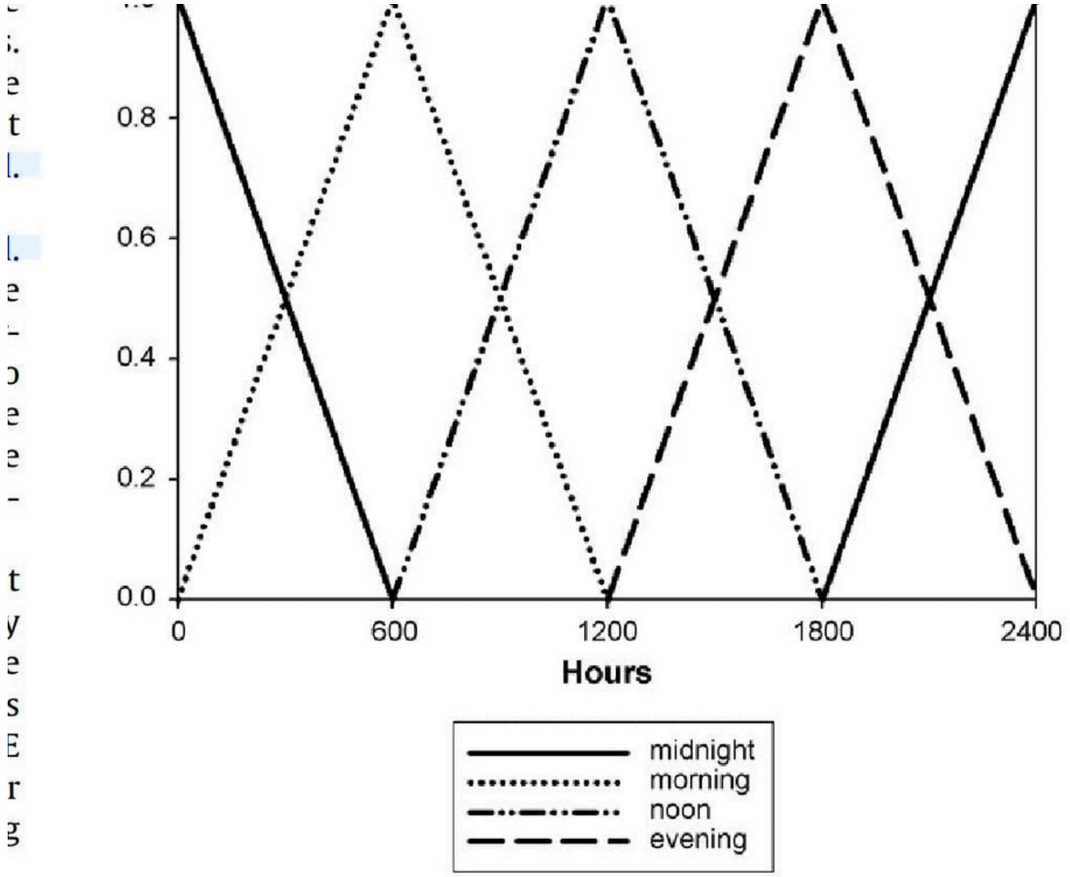
250 inputs were given to the model. The five main variables were: air temperature, solar radiation, wind speed, rainfall and relative humidity. Each value was scaled to [0.1, 0.9]. These were provided as:

- Value at the point of prediction
- Hourly rate of change at the point of prediction
- Values at 1h interval over the previous day
- Hourly rate of change at 1h interval over the previous day

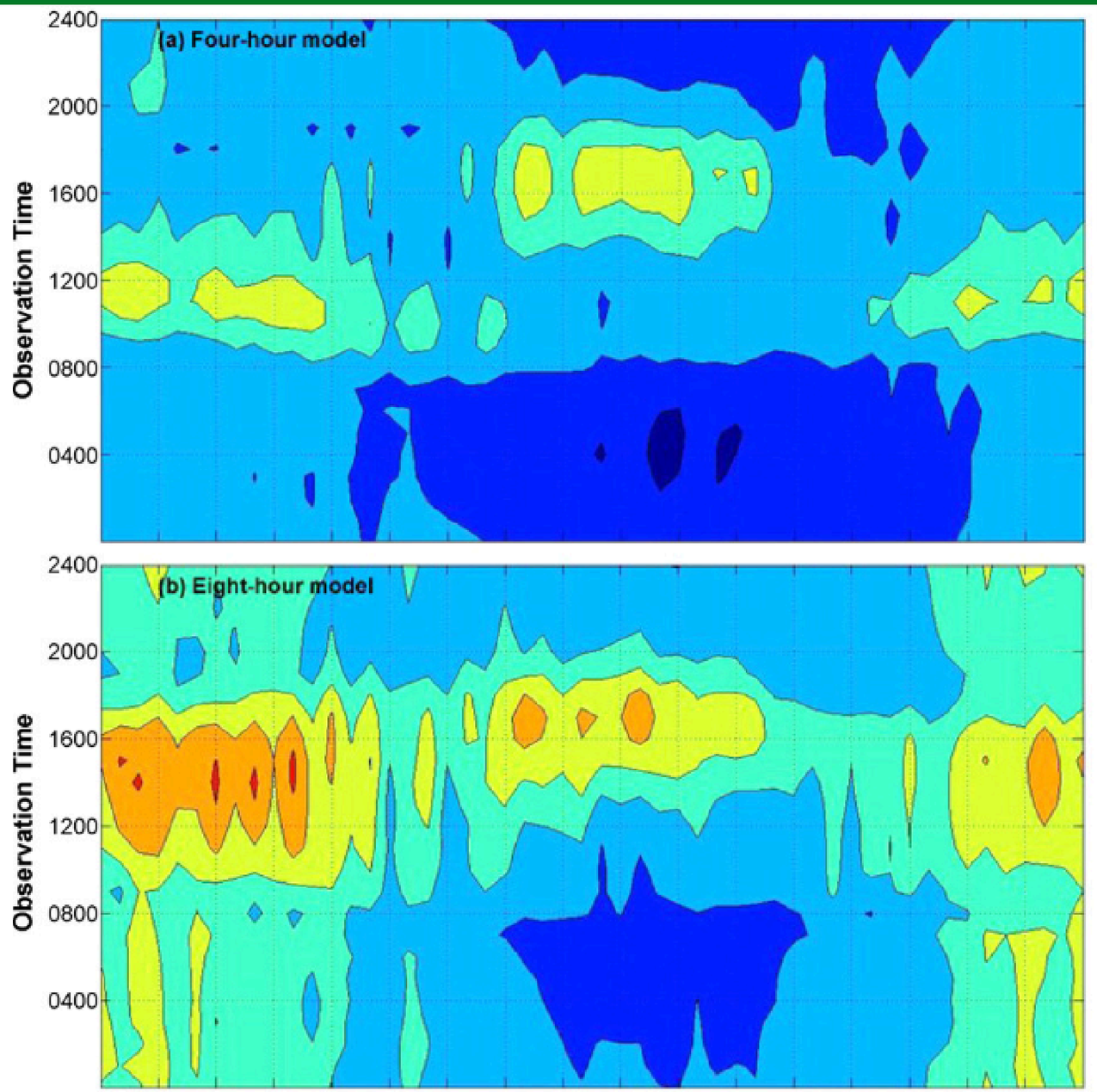
Models were trained separately on 300k values of winter data and randomly selected 300k values for year-round data.

Data & Variables

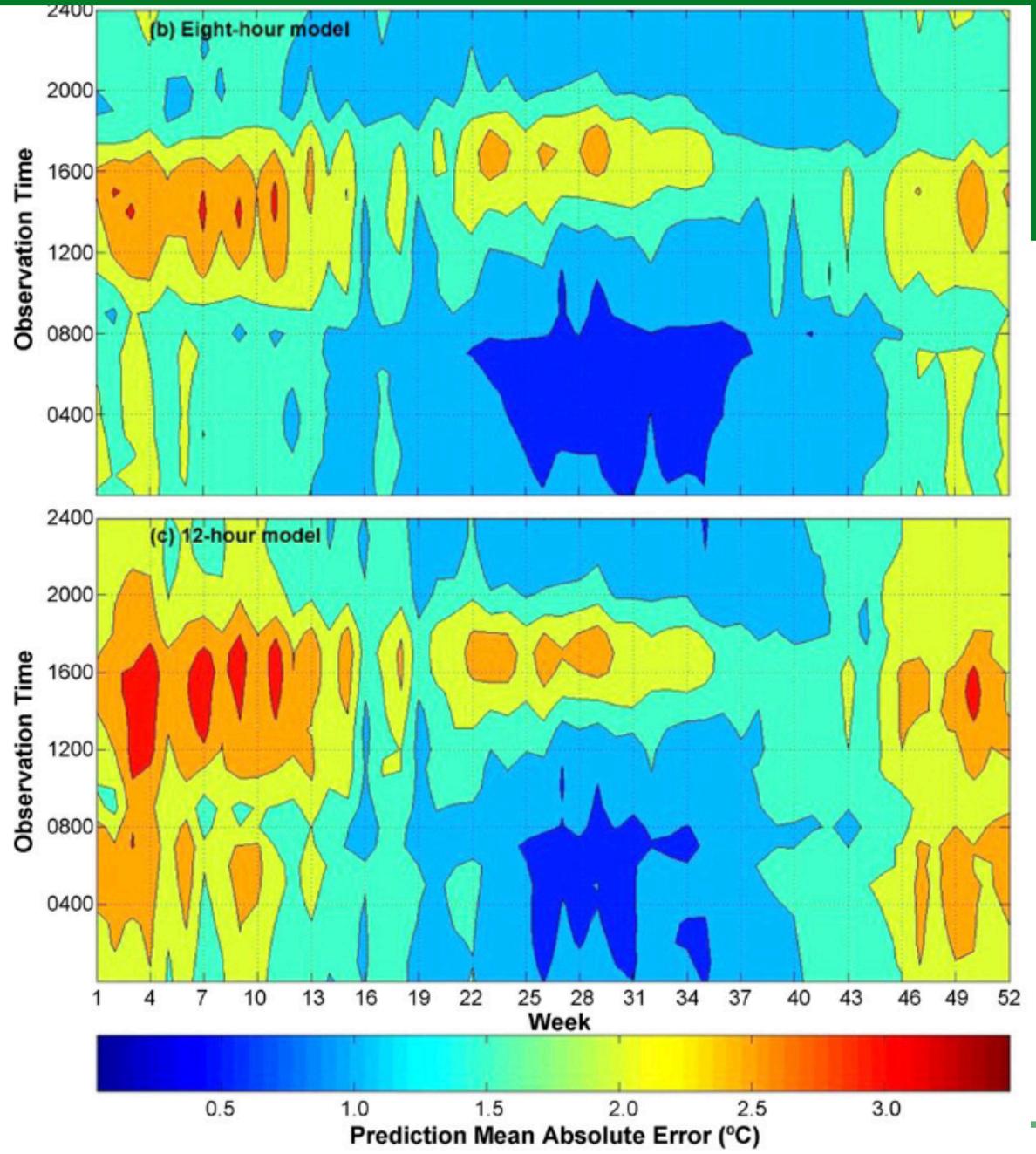
Time and seasonality were represented as four values for triangular fuzzy logic:



Findings


(a) Four-hour model
Observation Time
2400
2000
1600
1200
0800
0400
(b) Eight-hour model
Observation Time
2400
2000
1600
1200
0800
0400

Findings


(b) Eight-hour model
(c) 12-hour model
Observation Time
2400
2000
1600
1200
0800
0400
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52
Week
0.5 1.0 1.5 2.0 2.5 3.0
Prediction Mean Absolute Error (°C)

Thank you!