Recursive climate feature selection for regional weather prediction

Nathan Creaser, Kevin Donkers, Edward Pope, Samantha Adams, Anna-Louise Ellis, Andrew Cottrell, Jemma Davie

Context 1

Impact: China's North-East farming region accounts for ~30% of national maize production and ~5% of global maize production

Context: Previous work has produced a statistical model to predict maize yield using observed summer June-July-August (JJA) temperature and precipitation (Cottrell et al 2022 submitted)

The problem: This gives almost no lead time to respond to adverse scenarios. Our goal is to identify climate variables earlier in the year that can predict summer temperature and precipitation

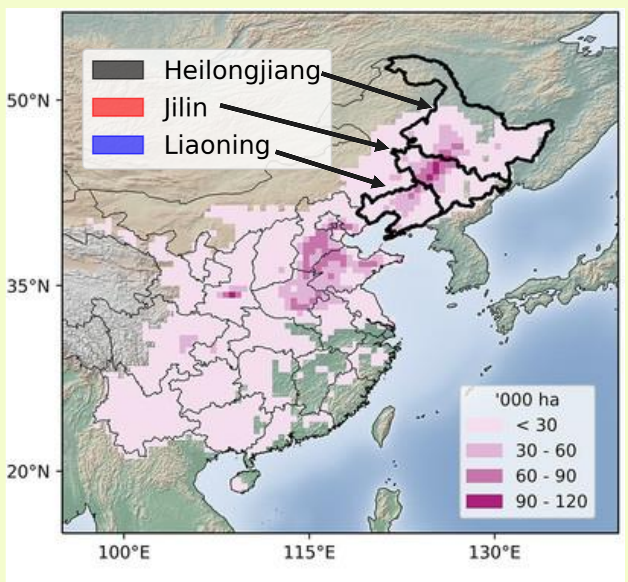


Figure 1: Maize harvest area over mainland China. For this study we focussed on the provinces of Liaoning, Jilin and Heilongjiang (Kent et al 2019)

Method 2

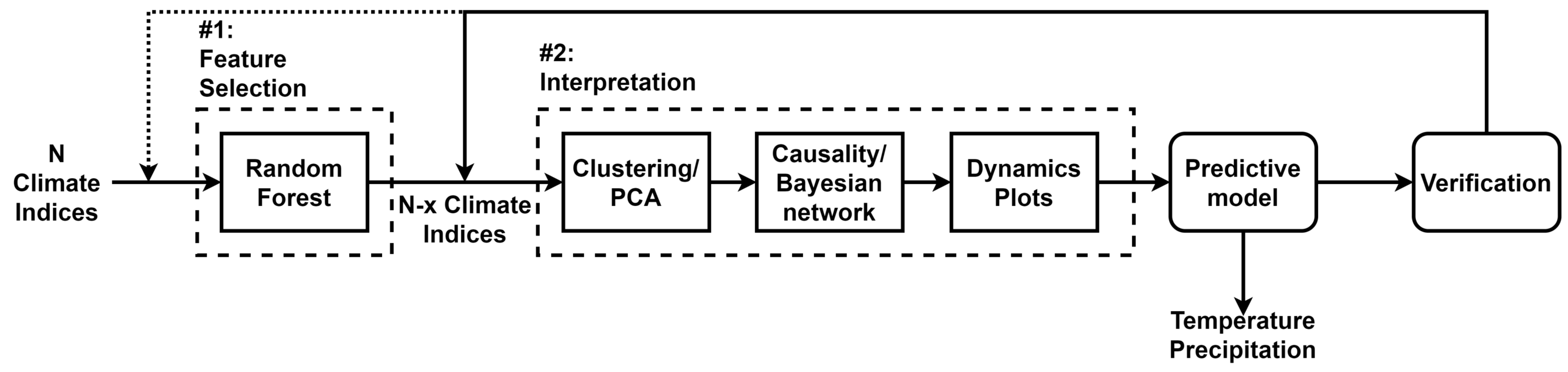


Figure 2: Recursive feature selection method for identifying important predictors of summer temperature and precipitation

Results 3

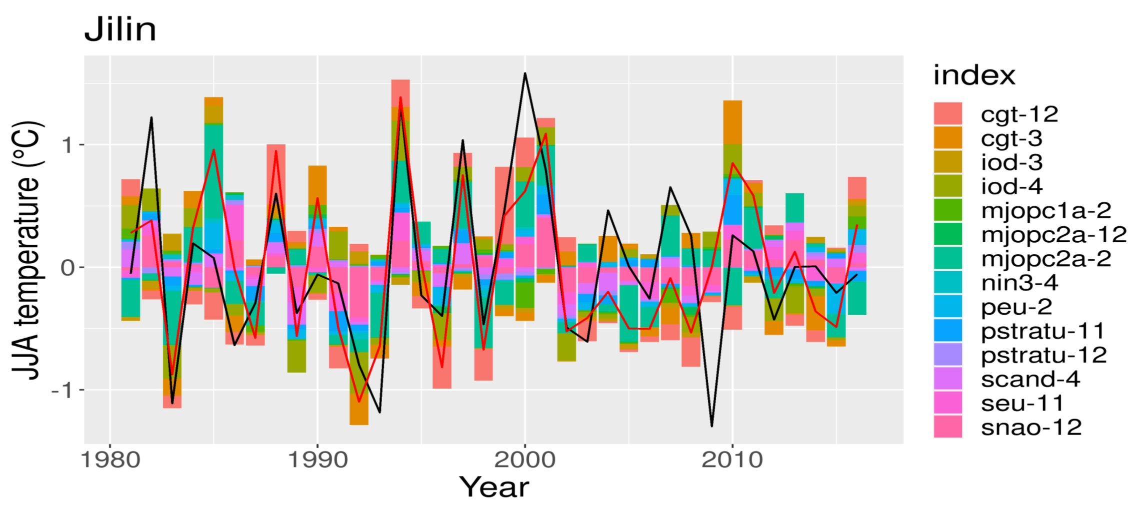


Figure 3: Example out-of-sample prediction of Jilin summer average temperature from a linear regression model using selected features.

Coloured stacked bars show the contribution from each predictor in each year from 1981-2016

	Temperature		Precipitation	
Liaoning	*0.81*	**0.75**	*0.75*	**0.76**
Jilin	*0.69*	**0.74**	*0.60*	**0.69**
Heilongjiang	*0.61*	**0.70**	*0.43*	**0.72**

Table 1: Pearson correlation between forecast and observed temperature and precipitation for all provinces.
Linear Regression Correlation | **Random Forest Regression Correlation**

Random forest regression performs more consistently than the linear regression because it can capture higher order relationships. This indicates interactions are more complex in Jilin and Heilongjiang.

Discovery 4

Random Forest

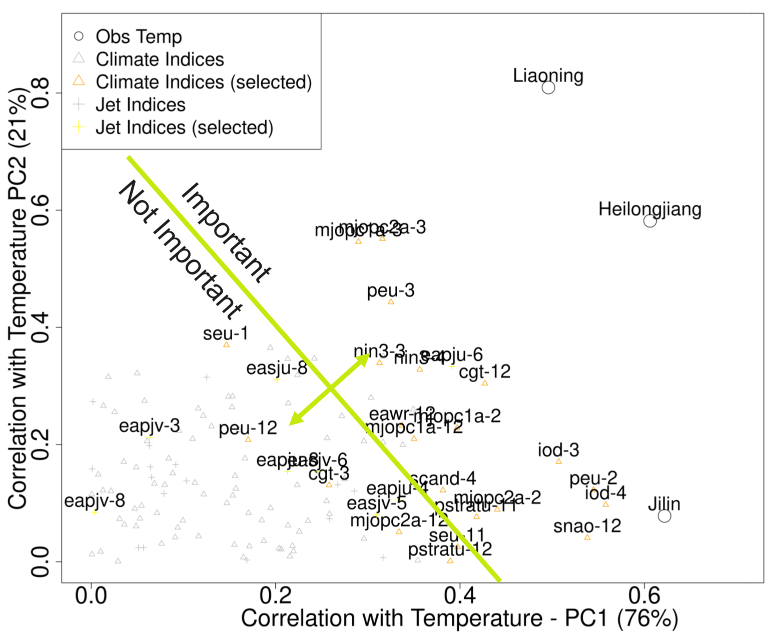


Figure 4: Shows the features selected by the random forest regression procedure

- Selected features are shown to be correlated with the first and second principle component of temperature across the three provinces.
- A straight line can be drawn approximately between selected and unselected features

Partial-Correlation Network

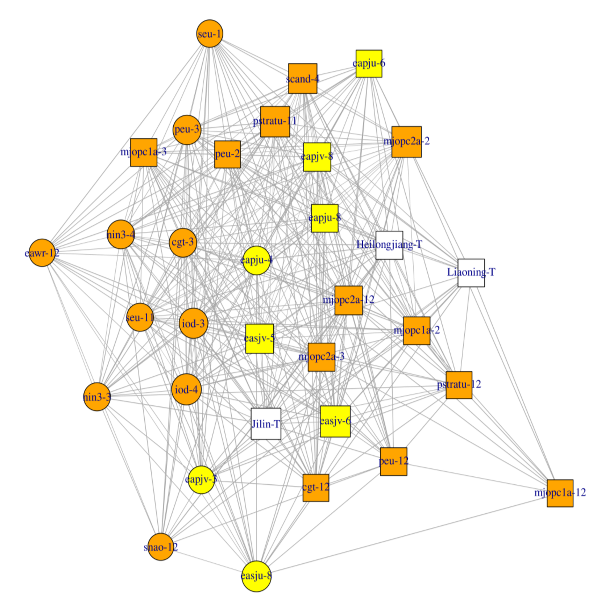
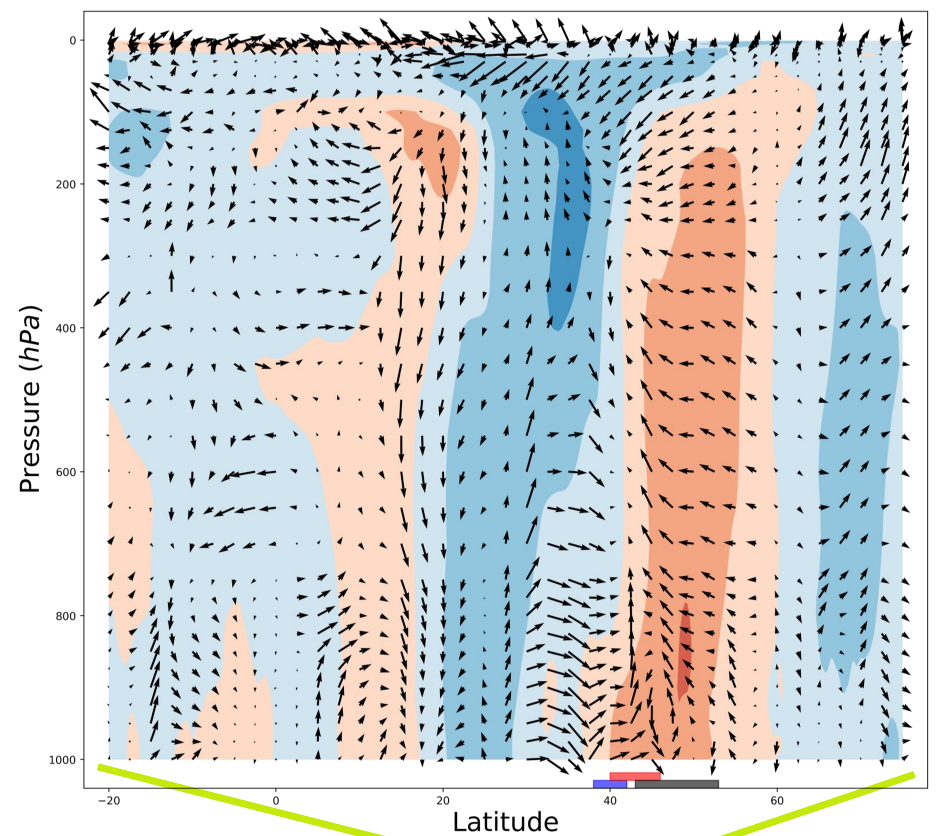


Figure 5: Partial-Correlation Network showing how predictors relate to summer temperature and each other

- White vertices show province temperature; orange vertices show selected climate features; yellow vertices show high-level jets – potential mediating influences.
- Distance between predictors signals partial strength of relationship
- Vertices show two interacting communities (circles and squares)

Interpretation 5



Dynamical relationships between March IOD and summer conditions.

Figure 6: Shading shows the correlation between March Indian Ocean Dipole (IOD) and averaged u wind between 120°E-135°E, the longitudinal range of the North-East farming region.
Quiver arrows show the regression of v and w winds averaged between 120°E-135°E onto the March IOD.
This highlights a strengthened southerly wind bringing warmer air to the south of the North-East farming region.

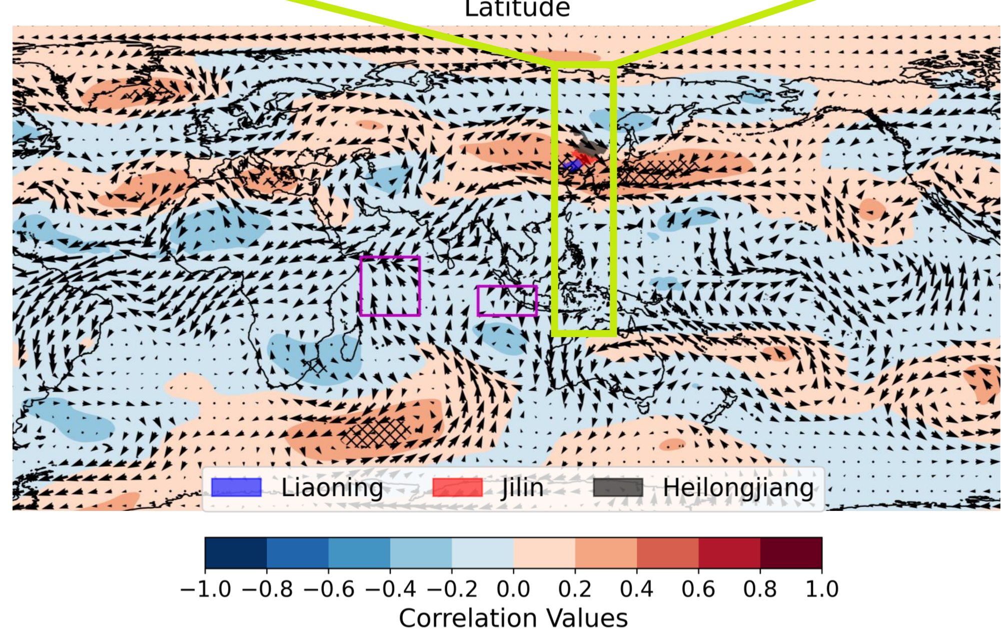


Figure 7: Shading shows the Pearson correlation between March IOD and 500 hPa Geopotential Height. Quiver arrows show the regression coefficients of u and v winds onto March IOD. Hatching indicates significance at 95% confidence level. Pink boxes define the IOD poles.

This highlights an association with a weakened subtropical jet and a strengthened subpolar jet, both of which are linked with hotter, drier conditions across North-East China.

Future Work 6

- Incorporating forecasts into maize yield climate service and assessing skill at different lead times
- Further assessment of causal relationships – what is ENSO's wider role?
- Demonstration of techniques in open-source R and Python notebooks

Capability 1: Discovery and attribution

Capability 2: Fusing simulation with data science

Capability 3: Uncertainty and trust

Capability 4: Data to decisions

Scan to learn more about these climate indices