Predicting the Weather (<u>without</u> physics)

Mark Fruman

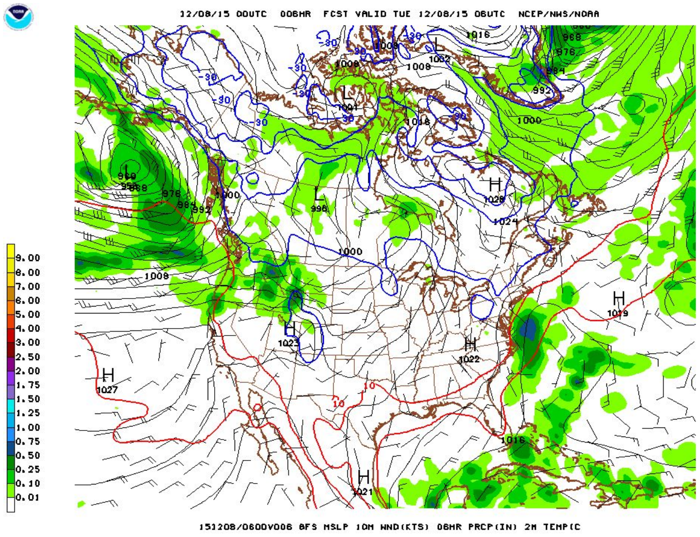
Capstone Project for Certificate in
Data Analytics, Big Data, and Predictive Analytics
Ryerson University, Toronto, ON, Canada

Advisor: Dr. Bora Çağlayan

github.com/majorgowan/wpwp

Ryerson University

Numerical Weather Prediction



http://www.emc.ncep.noaa.gov/GFS/

http://www.metoffice.gov.uk/

http://www.dwd.de/

$$\frac{Du}{Dt} - fv = -\frac{\partial \phi}{\partial x}$$

$$\frac{Dv}{Dt} + fu = -\frac{\partial \phi}{\partial y}$$

$$0 = -\frac{\partial \phi}{\partial p} - \frac{RT}{p}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial \omega}{\partial p} = 0$$

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} + \omega\left(\frac{\partial T}{\partial p} - \frac{RT}{pc_p}\right) = \frac{J}{c_p}$$

"Primitive Equations"

Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream.

— *Lewis Fry Richardson* —

Ryerson University

- Operational forecast models solve systems of coupled equations for **wind velocity**, **temperature**, **barometric pressure**, **density**, **gas concentrations** (e.g. water vapour), **etc** at each of **tens of millions** of points on a three-dimensional grid covering the forecast domain

- Data from worldwide network of **ground stations**, **satellites**, and **balloons** are used to initialize and constrain models (**“data assimilation”**)

- The equations are **sensitive to initial conditions** and highly **tuned** to produce realistic predictions (at the expense of physical rigour if necessary)

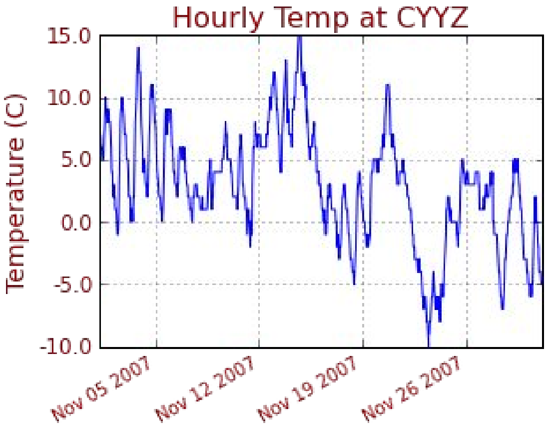
- **RMSE of between one and two degrees Celsius** for forecast of next-day daily maximum temperature (e.g. Silver 2012)

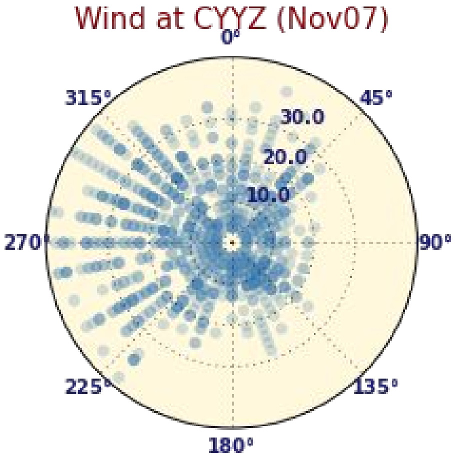
How well can plain linear regression do?

Ryerson University

Dataset

- Source: www.wunderground.com
- **Hourly** values of
  - Temperature, Dew Point Temperature
  - Relative Humidity
  - Sea Level Pressure
  - Visibility
  - Wind Speed and Direction, Wind Gust Speed
  - Precipitation
  - Events, Conditions





- Data from **29 stations**
- 8 years of data
  - **Training: 2005-2010**
  - **Testing: 2011-2012**
- Use daily summary statistics
  - Min/Max/Mean
  - Binary statistics

Ryerson University

Methodology

Multiple Linear Regression

- Assume **dependent variable** Y on the **forecast day** depends linearly on the values of the N **features** X_i on the **prediction day** (and/or earlier):

$$Y_{FD} = \beta_0 + \sum_{i=1}^{N} \beta_i X_i + \epsilon$$

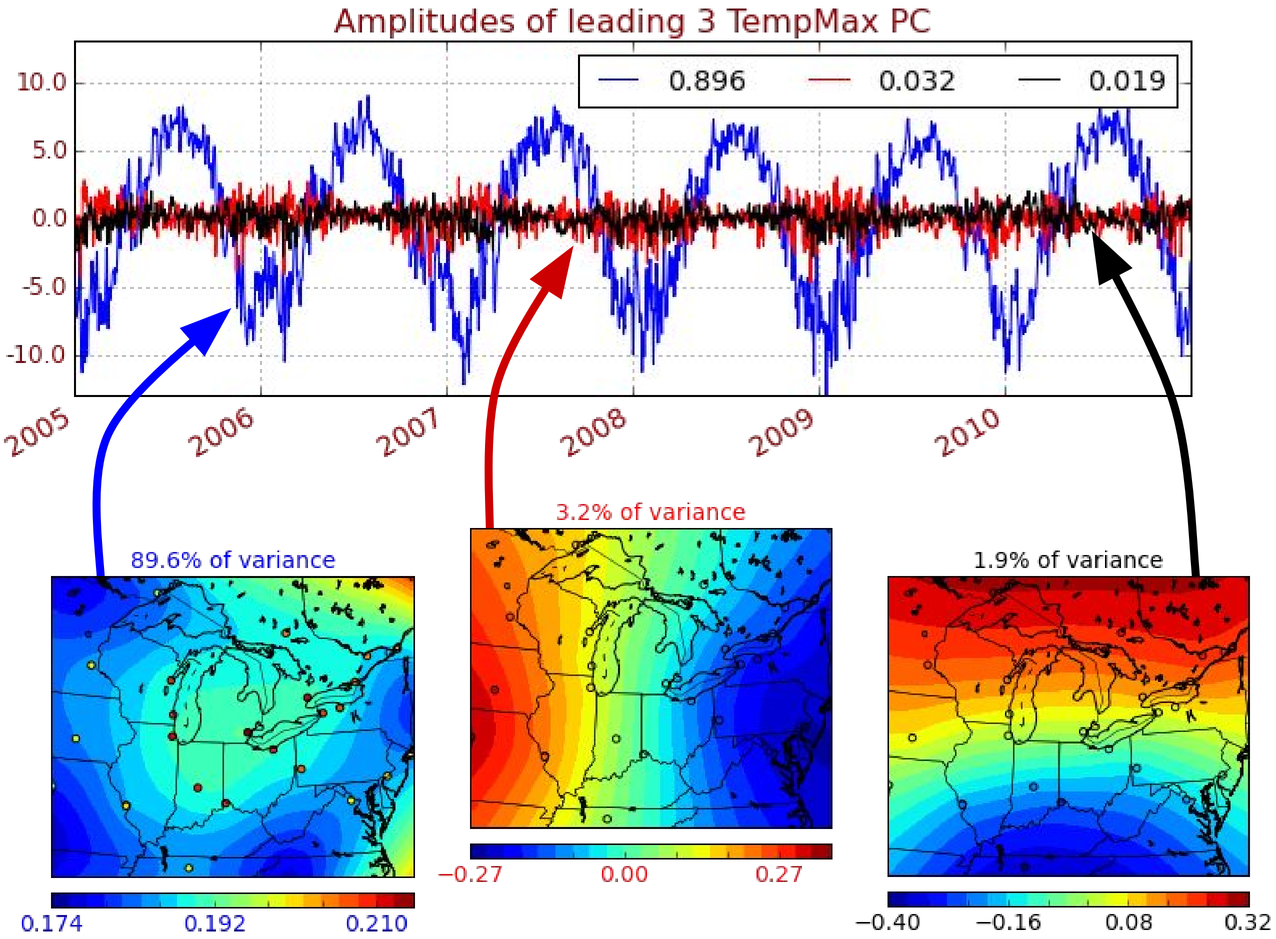
- Use values of X_i and Y in **training data** to find the $N+1$ coefficients β_0, β_i that minimize error ϵ

Variations

- X **includes only features** from the **prediction day** at **target station**
- **…** add features from days prior to the prediction day (**"Taylor"** model)
- **…** add features from **other stations** on prediction day and prior days
- Replace features with component blown towards target station (**"Advection"** model)
- Transform each feature at all stations into a **truncated set of principal components**
- Apply **k-means clustering** to training data and train a separate regression model for each cluster

Ryerson University

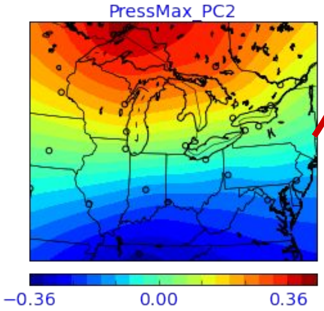
Principal Component Analysis

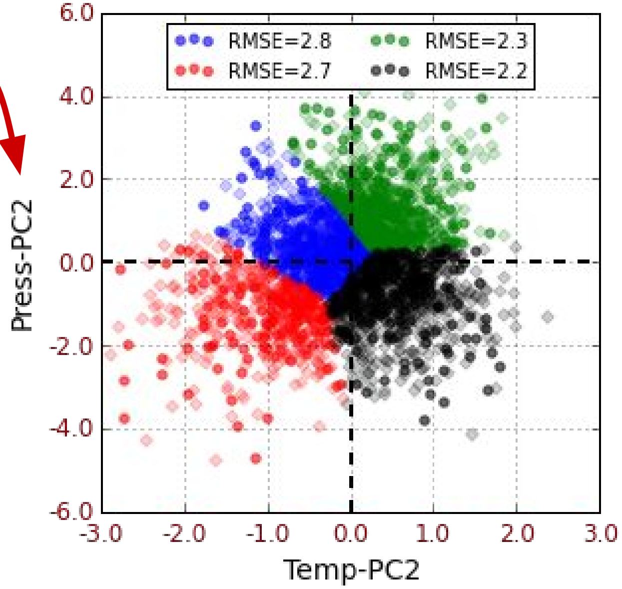


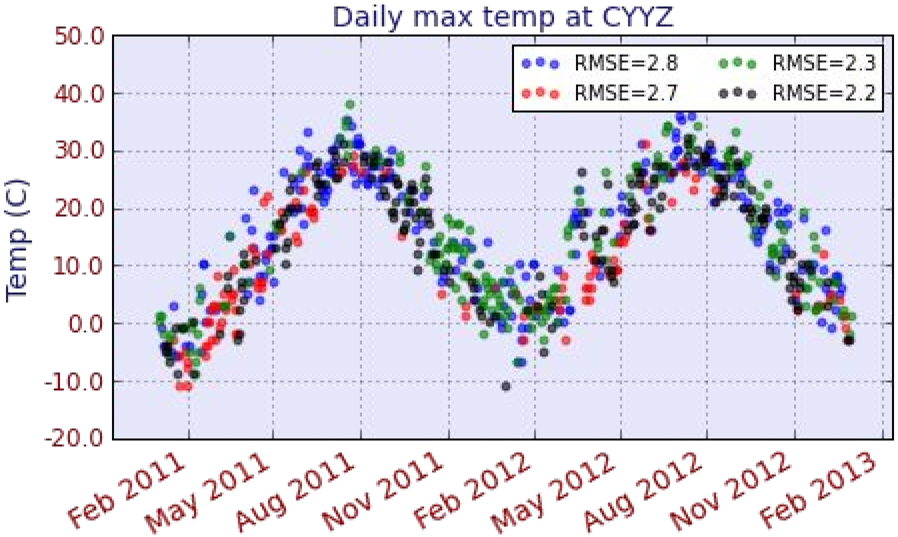
Ryerson University

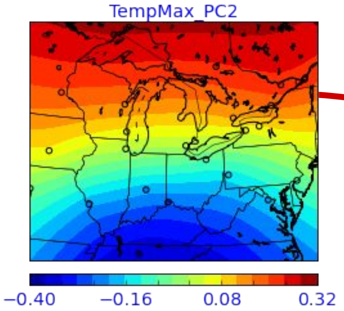
K-Means Clustering

- Compute **clusters** based on values of the third PC of `TempMax` and third PC of `PressMax` in training data
- Train **separate regression models for each cluster**
- Classify test points before making prediction using the appropriate model



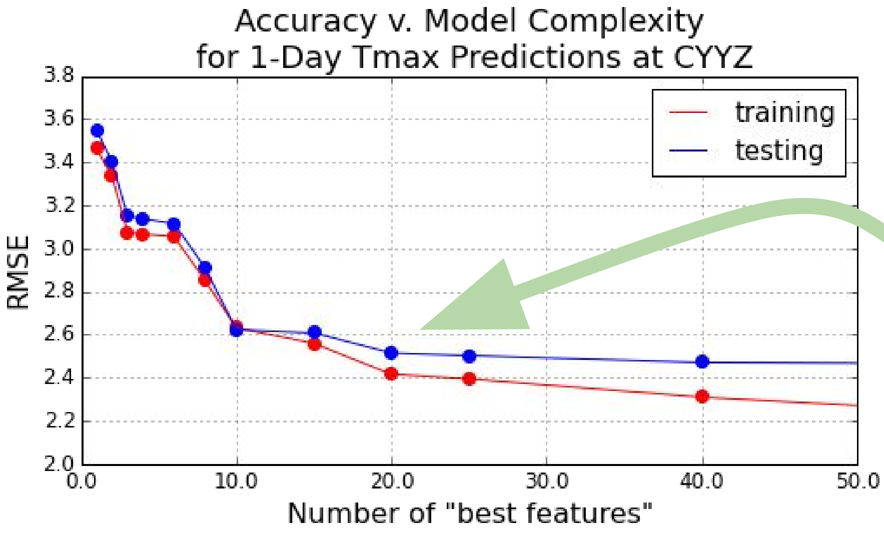






Ryerson University

Results: Conditions at CYYZ



- Apply **stepwise feature selection** starting with **2 days** of **5 PC** of each available variable
- For more than **20 features**, out-of-sample RMSE stops decreasing (**overfitting?**)

Features for best-20 model for TempMax at CYYZ

intercept	13.20		
TempMean_PC4	2.87	TempMin_PC0_D	0.61
TempMean_PC0	2.71	TempMin_PC1	0.58
TempMax_PC2	1.22	TempMax_PC3	-0.56
TempMin_PC4	1.19	PressMin_PC2_D	-0.41
TempMax_PC4	-1.17	dailyPressRange_PC3	-0.38
TempMin_PC0	-1.01	TempMax_PC0	0.38
TempMin_PC3	0.98	WindMeanY_PC1	-0.35
TempMean_PC4_D	-0.69	dailyTempRange_PC4	-0.35
PressMax_PC2_D	0.66	dailyTempRange_PC0_D	0.33
PressMean_PC2_D	-0.63	isMorningMinTemp_PC3	-0.10

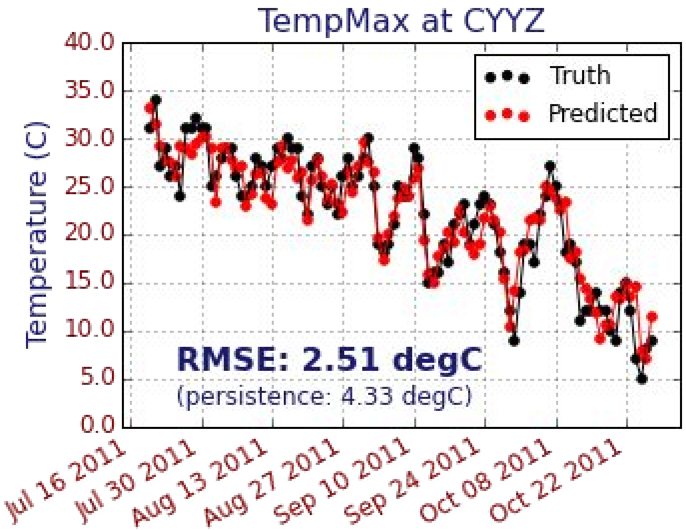
scaled regression coefficients

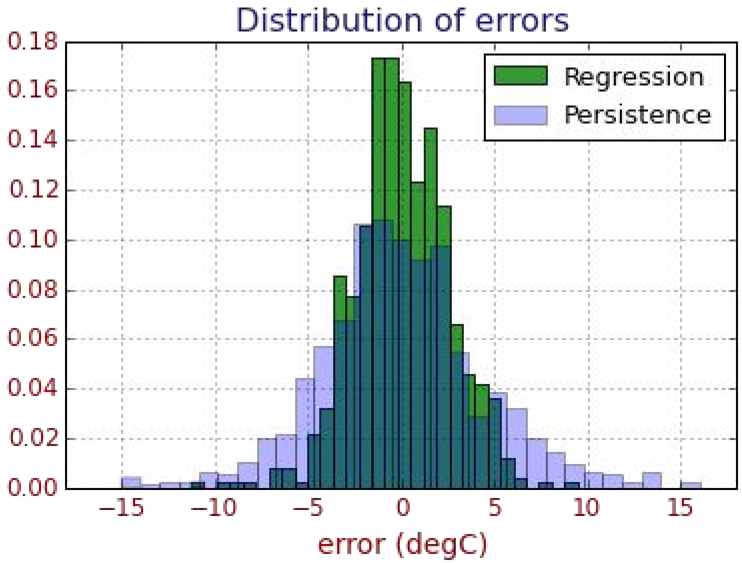
N.B. quite a bit of multicollinearity

Ryerson University

Results: Conditions at CYYZ

Event (num)	RMS change	RMSE (% better)
0-sigma (460)	**2.39 C**	**2.11 C** (64.1%)
1-sigma (151)	**6.13 C**	**3.07 C** (96.7%)
2-sigma (31)	**10.15 C**	**3.80 C** (100%)
3-sigma (11)	**14.05 C**	**5.89 C** (100%)



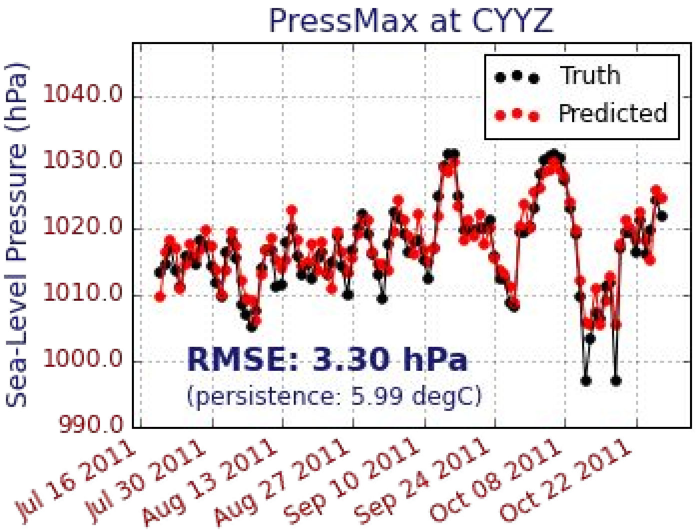


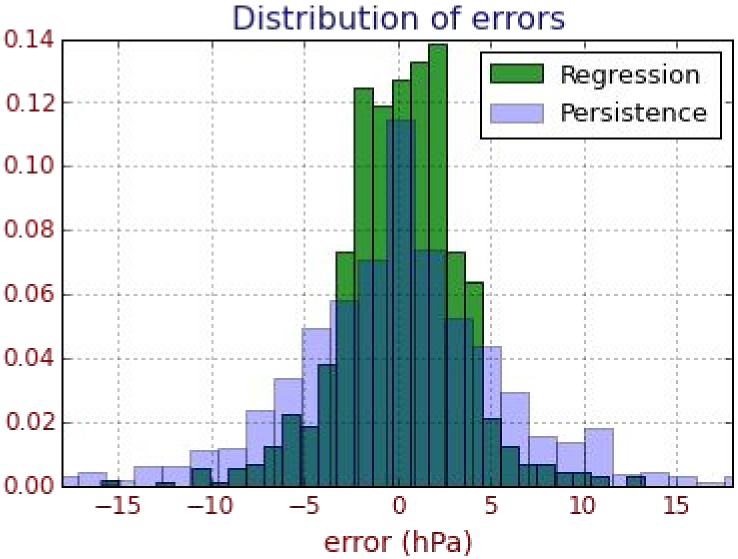
acc. 77%	Pred Colder	Pred Warmer
Colder	**276**	**108**
Warmer	**58**	**286**

Ryerson University

Results: Conditions at CYYZ

Event (num)	RMS change	RMSE (% better)
0-sigma (427)	**2.18 hPa**	**2.47 hPa** (50.4%)
1-sigma (189)	**6.10 hPa**	**3.47 hPa** (94.7%)
2-sigma (68)	**10.64 hPa**	**4.47 hPa** (100%)
3-sigma (19)	**14.94 hPa**	**5.89 hPa** (100%)
4+-sigma (13)	**19.80 hPa**	**8.55 hPa** (100%)

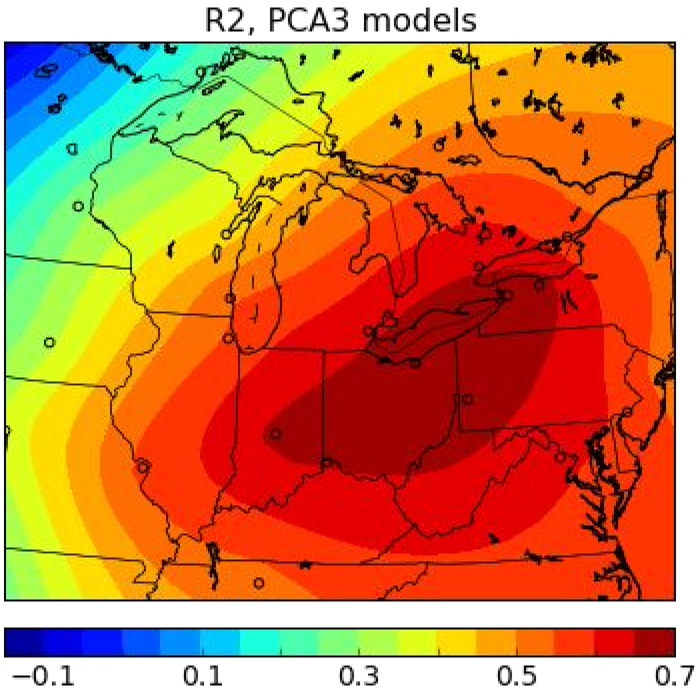


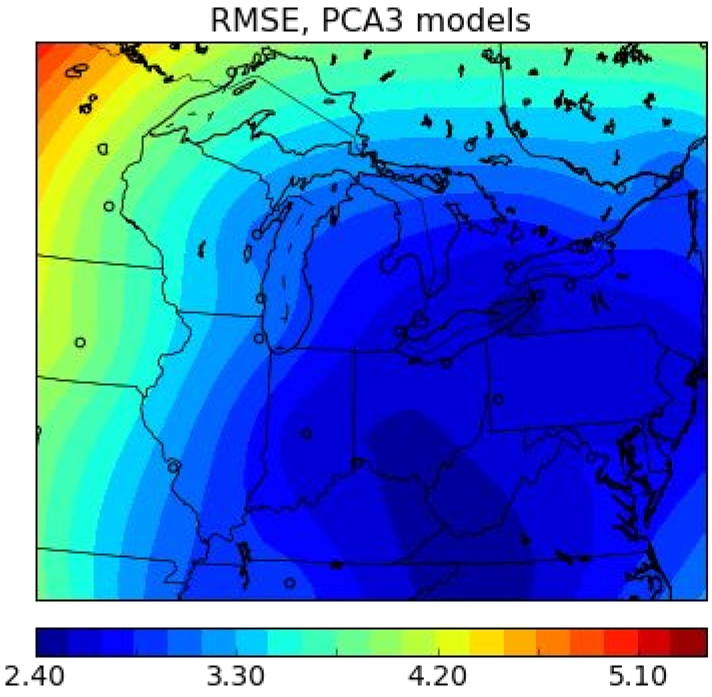


acc. 81%	Pred Fall	Pred Rise
Fall	**292**	**75**
Rise	**62**	**299**

Ryerson University

Results: Predictability vs. Station Location





- R^2 relative to **persistence forecast** for `TempMax` predictions with 3-PC regression model
- RMSE for `TempMax` predictions with 3-PC regression model

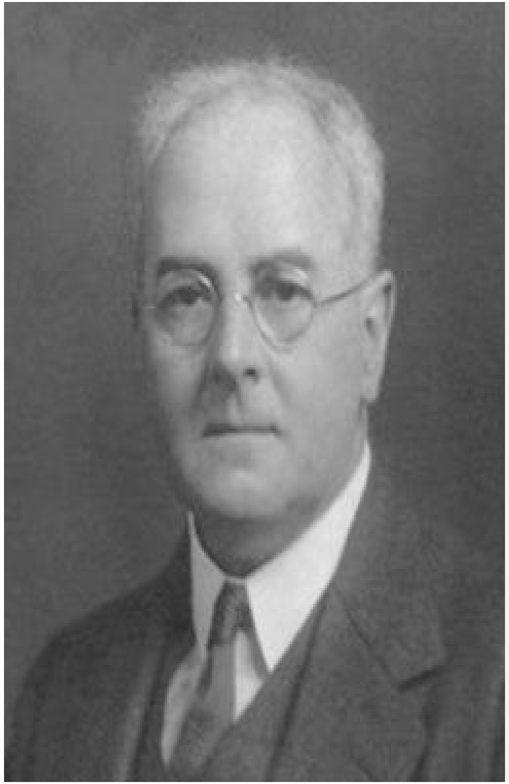
Ryerson University

Conclusion

- Operational models report **RMSE of between one and two degrees Celsius** for forecast of next-day daily max. temperature (cf. *“3-degree guarantee”*)

- Best regression model (**20 PC features**) has RMSE for Toronto of **2.51 degC**

- Software implemented in Python:
  - ✓ Automatically **harvests** and **archives** data in **JSON format**
  - ✓ Computes and stores daily **summary statistics** in **CSV format**
  - ✓ Predicts future value of **any daily summary statistic** for **any station** using **any combination of features** from **any number of stations** with or without using **PCA** and **k-means clustering**
  - ✓ Can be applied to any set of multivariate time-series data
  - ✓ Available at `github.com/majorgowan/wpwp`



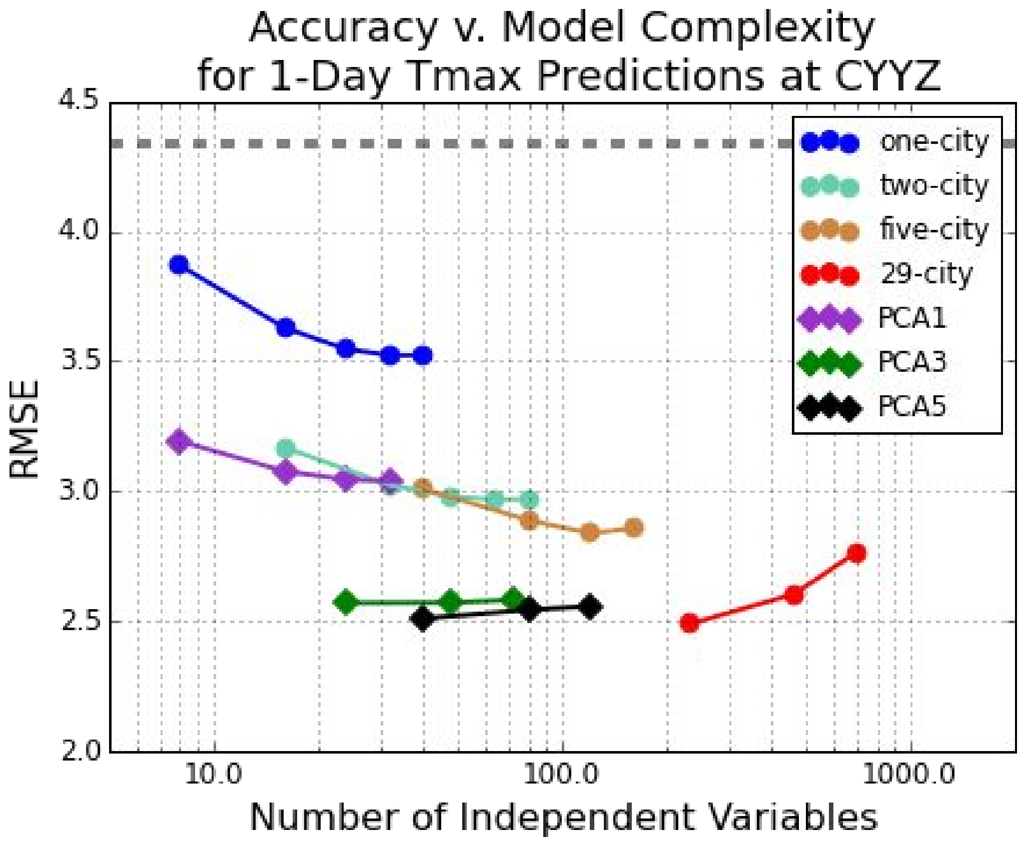
Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream.

— Lewis Fry Richardson —

(1922)

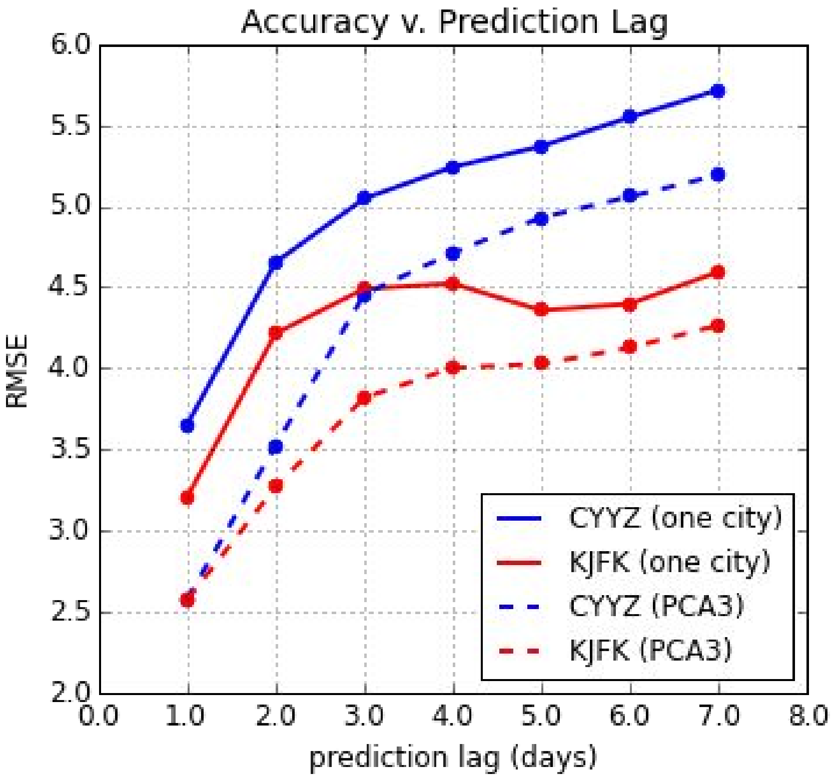
Ryerson University

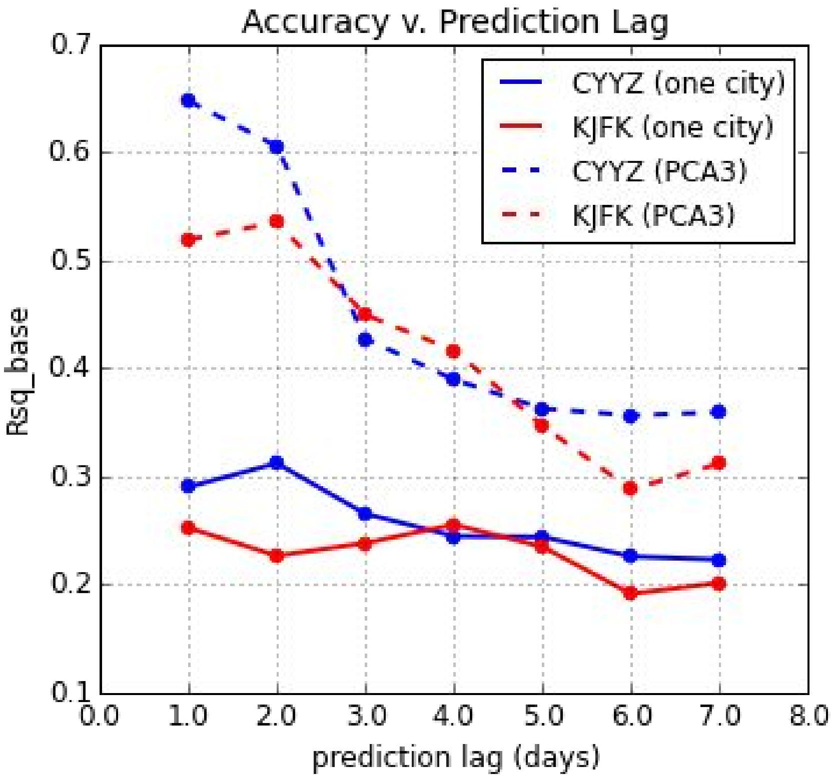
Model Accuracy vs Number of Features



Ryerson University

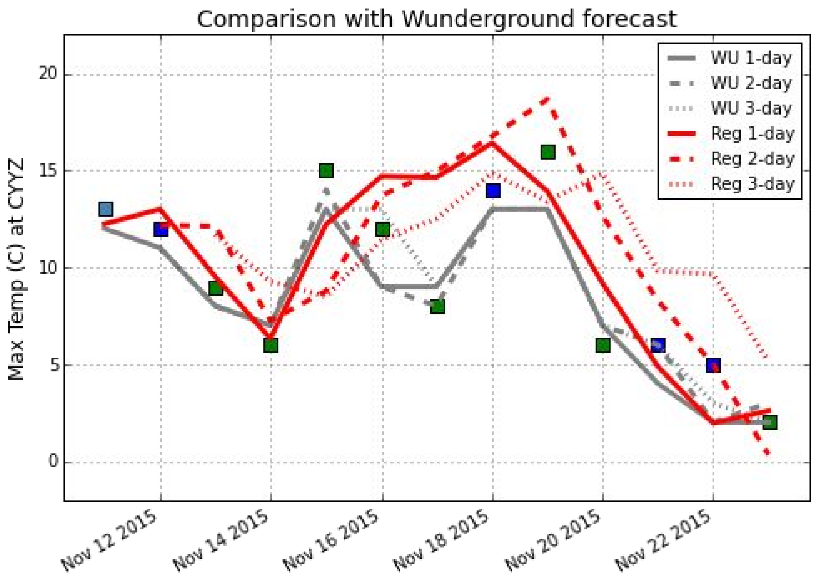
Model Accuracy vs Lead Time

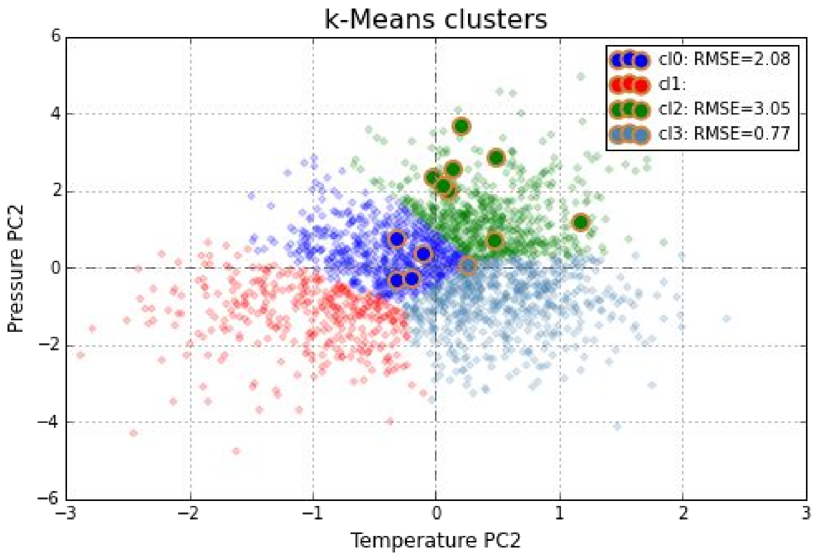




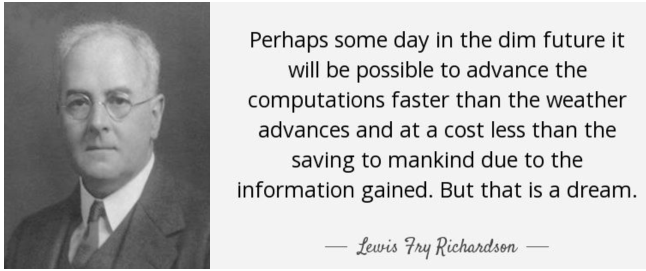
Ryerson University

November 11-23, 2015





Ryerson University



$$Y_{FD} = \beta_0 + \sum_{i=1}^{N} \beta_i X_i + \epsilon$$

β_0, β_i

ϵ

1. problem definition
2. techniques used in the literature
3. dataset description
4. methodology
5. results and discussion
6. results and discussion
7. results and discussion
8. conclusion

$$\frac{Du}{Dt} - fv = -\frac{\partial \phi}{\partial x}$$

$$\frac{Dv}{Dt} + fu = -\frac{\partial \phi}{\partial y}$$

$$0 = -\frac{\partial \phi}{\partial p} - \frac{RT}{p}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial \omega}{\partial p} = 0$$

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} + \omega\left(\frac{\partial T}{\partial p} - \frac{RT}{pc_p}\right) = \frac{J}{c_p}$$