Improving Subseasonal Forecasting in the Western U.S.

Paulo Orenstein

March 22, 2019

Photo credit: IIP Photo Archive

Joint work with Jessica Hwang, Lester Mackey, Judah Cohen, Karl Pfeiffer

Goals

- Bring awareness to subseasonal forecasting, an important problem for water management and weather extremes

Goals

- Bring awareness to subseasonal forecasting, an important problem for water management and weather extremes
- Introduce an example of a crowdsourced, social good project

Goals

- Bring awareness to subseasonal forecasting, an important problem for water management and weather extremes
- Introduce an example of a crowdsourced, social good project
- Present the `SubseasonalRodeo Dataset`

Goals

- Bring awareness to subseasonal forecasting, an important problem for water management and weather extremes
- Introduce an example of a crowdsourced, social good project
- Present the `SubseasonalRodeo Dataset`
- Discuss effective machine learning methods for the problem

Goals

- Bring awareness to subseasonal forecasting, an important problem for water management and weather extremes
- Introduce an example of a crowdsourced, social good project
- Present the `SubseasonalRodeo Dataset`
- Discuss effective machine learning methods for the problem
  - multitask model selection
  - weighted locally linear regression
  - ensembling

Goals

- Bring awareness to subseasonal forecasting, an important problem for water management and weather extremes
- Introduce an example of a crowdsourced, social good project
- Present the `SubseasonalRodeo Dataset`
- Discuss effective machine learning methods for the problem
  - multitask model selection
  - weighted locally linear regression
  - ensembling
- Encourage you to improve on our results!

Motivation

- Long-term weather prediction (> 2 months): hopeless, use historical climate

Motivation

- Long-term weather prediction ($>$ 2 months): hopeless, use historical climate
- Short-term weather prediction ($<$ 2 weeks): accurate predictions possible using physics-based models

Motivation

- Long-term weather prediction (> 2 months): hopeless, use historical climate
- Short-term weather prediction (< 2 weeks): accurate predictions possible using physics-based models
- Medium-term (*subseasonal*) weather prediction: physics-based models are no longer accurate

Motivation

- Long-term weather prediction (> 2 months): hopeless, use historical climate
- Short-term weather prediction (< 2 weeks): accurate predictions possible using physics-based models
- Medium-term (*subseasonal*) weather prediction: physics-based models are no longer accurate
- Subseasonal forecasts are important

Motivation

- Long-term weather prediction (> 2 months): hopeless, use historical climate
- Short-term weather prediction (< 2 weeks): accurate predictions possible using physics-based models
- Medium-term (*subseasonal*) weather prediction: physics-based models are no longer accurate
- Subseasonal forecasts are important
  - allocate water resources
  - manage wildfires
  - prepare for droughts, floods and other weather extremes
  - crop planting, irrigation scheduling, and fertilizer application

Motivation

- Long-term weather prediction (> 2 months): hopeless, use historical climate
- Short-term weather prediction (< 2 weeks): accurate predictions possible using physics-based models
- Medium-term (*subseasonal*) weather prediction: physics-based models are no longer accurate
- Subseasonal forecasts are important
  - allocate water resources
  - manage wildfires
  - prepare for droughts, floods and other weather extremes
  - crop planting, irrigation scheduling, and fertilizer application
- Can statistical/ML/non-physics models extend the forecast horizon beyond short-term prediction?

"During the past eight years, every state in the Western United States has experienced drought that has affected the economy both locally and nationally through impacts to agricultural production, water supply, and energy."

David Raff, USBR

Forecasting systems in use now

- CFSv2 (Climate Forecasting System, version 2): operational forecasting system for the US, physics-based model representing "coupled atmosphere-ocean-land surface-sea ice system"

Forecasting systems in use now

- CFSv2 (Climate Forecasting System, version 2): operational forecasting system for the US, physics-based model representing “coupled atmosphere-ocean-land surface-sea ice system”

- NMME (North American Model Ensemble): ensemble of CFSv2 and about 10 other physics-based models from various North American modeling centers

Forecasting systems in use now

- CFSv2 (Climate Forecasting System, version 2): operational forecasting system for the US, physics-based model representing "coupled atmosphere-ocean-land surface-sea ice system"
- NMME (North American Model Ensemble): ensemble of CFSv2 and about 10 other physics-based models from various North American modeling centers
- Both are examples of Numerical Weather Prediction models

Forecasting systems in use now

- CFSv2 (Climate Forecasting System, version 2): operational forecasting system for the US, physics-based model representing "coupled atmosphere-ocean-land surface-sea ice system"

- NMME (North American Model Ensemble): ensemble of CFSv2 and about 10 other physics-based models from various North American modeling centers

- Both are examples of Numerical Weather Prediction models
  - simulate future weather using partial differential equations and supercomputers
  - initialized many times with current weather conditions; use the average of predictions
  - initial error doubles every 5 days

Forecasting systems in use now

- CFSv2 (Climate Forecasting System, version 2): operational forecasting system for the US, physics-based model representing "coupled atmosphere-ocean-land surface-sea ice system"
- NMME (North American Model Ensemble): ensemble of CFSv2 and about 10 other physics-based models from various North American modeling centers
- Both are examples of Numerical Weather Prediction models
  - simulate future weather using partial differential equations and supercomputers
  - initialized many times with current weather conditions; use the average of predictions
  - initial error doubles every 5 days
- Can we do better?

Subseasonal Climate Forecast Rodeo

- Year-long, real-time forecasting competition sponsored by US Bureau of Reclamation and NOAA

Subseasonal Climate Forecast Rodeo

- Year-long, real-time forecasting competition sponsored by US Bureau of Reclamation and NOAA
- Four categories

Subseasonal Climate Forecast Rodeo

- Year-long, real-time forecasting competition sponsored by US Bureau of Reclamation and NOAA
- Four categories
  - two variables: two-week average temperature and two-week accumulated precipitation
  - two forecasting horizons: 3-4 weeks out and 5-6 weeks out

Subseasonal Climate Forecast Rodeo

- Year-long, real-time forecasting competition sponsored by US Bureau of Reclamation and NOAA
- Four categories
  - two variables: two-week average temperature and two-week accumulated precipitation
  - two forecasting horizons: 3-4 weeks out and 5-6 weeks out
- Submission frequency: every two weeks

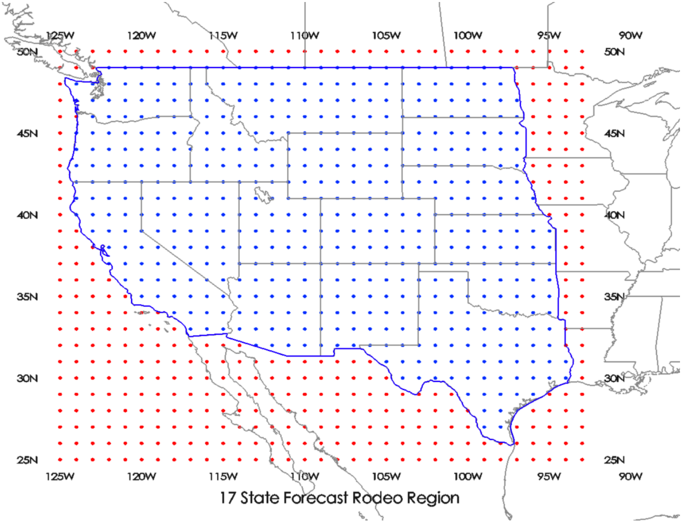
Subseasonal Climate Forecast Rodeo

- Year-long, real-time forecasting competition sponsored by US Bureau of Reclamation and NOAA
- Four categories
  - two variables: two-week average temperature and two-week accumulated precipitation
  - two forecasting horizons: 3-4 weeks out and 5-6 weeks out
- Submission frequency: every two weeks
  - first submission: April 18, 2017
  - last submission: April 3, 2018

Subseasonal Climate Forecast Rodeo

- Year-long, real-time forecasting competition sponsored by US Bureau of Reclamation and NOAA
- Four categories
  - two variables: two-week average temperature and two-week accumulated precipitation
  - two forecasting horizons: 3-4 weeks out and 5-6 weeks out
- Submission frequency: every two weeks
  - first submission: April 18, 2017
  - last submission: April 3, 2018
- Region: 17 states in western US, $G = 514$ grid points

Forecast Rodeo region



Contest scoring/objective

- For the two-week period beginning on t

Contest scoring/objective

- For the two-week period beginning on t
  - observed average temperature or total precipitation: $\mathbf{y}_t \in \mathbf{R}^G$
  - *climatology* for a month-day combination d:

$$\mathbf{c}_d = \frac{1}{30} \sum_{\substack{t\,:\,\mathtt{monthday}(t)=d,\\ 1981 \le \mathtt{year}(t) \le 2010}} \mathbf{y}_t$$

    the long-term average over 1981-2010 for the month-day d
  - observed *anomaly*: $\mathbf{a}_t = \mathbf{y}_t - \mathbf{c}_{\mathtt{monthday}(t)}$

Contest scoring/objective

- For the two-week period beginning on t
  - observed average temperature or total precipitation: $\mathbf{y}_t \in \mathbf{R}^G$
  - *climatology* for a month-day combination d:

$$\mathbf{c}_d = \frac{1}{30} \sum_{\substack{t\,:\,\texttt{monthday}(t)=d,\\ 1981 \leq \texttt{year}(t) \leq 2010}} \mathbf{y}_t$$

    the long-term average over 1981-2010 for the month-day d
  - observed *anomaly*: $\mathbf{a}_t = \mathbf{y}_t - \mathbf{c}_{\texttt{monthday}(t)}$

- Given a forecast $\hat{\mathbf{y}}_t$, or equivalently, forecast anomalies $\hat{\mathbf{a}}_t = \hat{\mathbf{y}}_t - \mathbf{c}_{\texttt{monthday}(t)}$, the cosine similarity or *skill* is

$$\text{skill}(\hat{\mathbf{a}}_t, \mathbf{a}_t) = \cos(\hat{\mathbf{a}}_t, \mathbf{a}_t) = \frac{\langle \hat{\mathbf{a}}_t, \mathbf{a}_t \rangle}{\|\hat{\mathbf{a}}_t\|_2 \|\mathbf{a}_t\|_2}$$

  Highest average skill over the contest period = winner

Contest scoring/objective

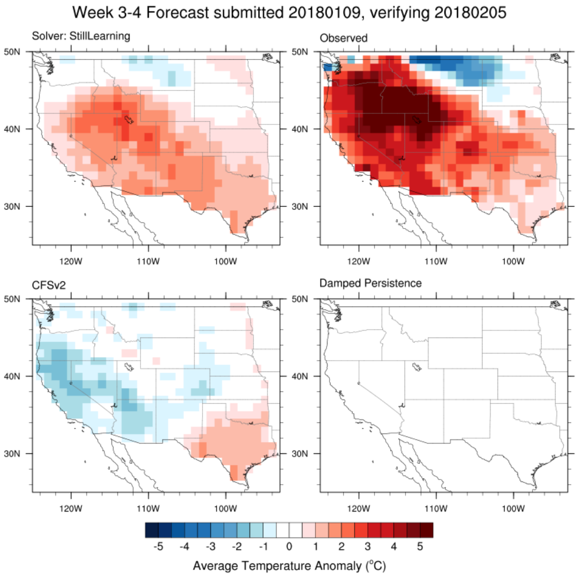
- For the two-week period beginning on t
  - observed average temperature or total precipitation: $\mathbf{y}_t \in \mathbf{R}^G$
  - *climatology* for a month-day combination d:

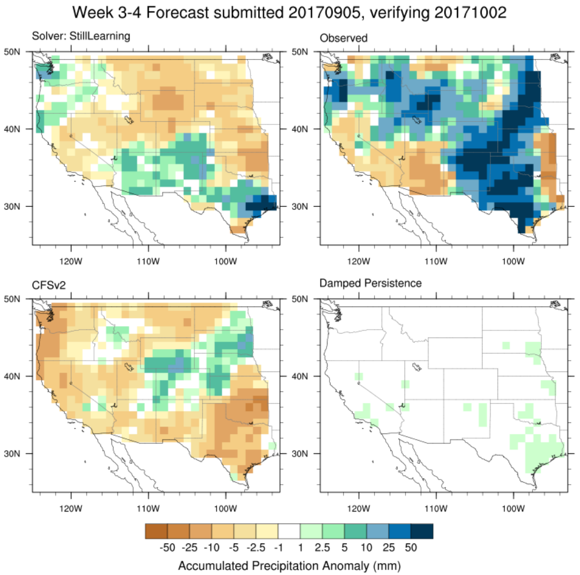
$$\mathbf{c}_d = \frac{1}{30} \sum_{\substack{t\,:\,\mathtt{monthday}(t)=d,\\ 1981 \le \mathtt{year}(t) \le 2010}} \mathbf{y}_t$$

    the long-term average over 1981-2010 for the month-day d
  - observed *anomaly*: $\mathbf{a}_t = \mathbf{y}_t - \mathbf{c}_{\mathtt{monthday}(t)}$
- Given a forecast $\hat{\mathbf{y}}_t$, or equivalently, forecast anomalies $\hat{\mathbf{a}}_t = \hat{\mathbf{y}}_t - \mathbf{c}_{\mathtt{monthday}(t)}$, the cosine similarity or *skill* is

$$\text{skill}(\hat{\mathbf{a}}_t, \mathbf{a}_t) = \cos(\hat{\mathbf{a}}_t, \mathbf{a}_t) = \frac{\langle \hat{\mathbf{a}}_t, \mathbf{a}_t \rangle}{\|\hat{\mathbf{a}}_t\|_2 \|\mathbf{a}_t\|_2}$$

  Highest average skill over the contest period = winner
- Benchmarks: debiased CFSv2 and "damped persistence"


Week 3-4 Forecast submitted 20180109, verifying 20180205
Solver: StillLearning
Observed
CFSv2
Damped Persistence
50N
40N
30N
120W
110W
100W
-5 -4 -3 -2 -1 0 1 2 3 4 5
Average Temperature Anomaly (°C)


Week 3-4 Forecast submitted 20170905, verifying 20171002
Solver: StillLearning
Observed
CFSv2
Damped Persistence
50N
40N
30N
120W
110W
100W
-50 -25 -10 -5 -2.5 -1 1 2.5 5 10 25 50
Accumulated Precipitation Anomaly (mm)

Our dataset

- No data provided!

Our dataset

- No data provided!
- Gathered historical data on various weather variables, 1980 to present

Our dataset

- No data provided!
- Gathered historical data on various weather variables, 1980 to present
  - temperature
  - precipitation
  - sea surface temperature
  - sea ice concentration
  - multivariate El Niño / Southern Oscillation index
  - Madden-Julian oscillation
  - relative humidity
  - pressure
  - geopotential height
  - historical NMME forecasts

Our dataset

- No data provided!
- Gathered historical data on various weather variables, 1980 to present
  - temperature
  - precipitation
  - sea surface temperature
  - sea ice concentration
  - multivariate El Niño / Southern Oscillation index
  - Madden-Julian oscillation
  - relative humidity
  - pressure
  - geopotential height
  - historical NMME forecasts
- Some are spatiotemporal, some are only temporal

Our dataset

- This was an enormous amount of work (seriously)

Our dataset

- This was an enormous amount of work (seriously)
- Postprocessing and transformation

Our dataset

- This was an enormous amount of work (seriously)
- Postprocessing and transformation
  - aggregated to two-week averages or sums
  - PCAed some variables to reduce dimensionality down to 3 principal components
  - chose 1-3 fixed lags for each variable according to forecast horizon, data availability

Our dataset

- This was an enormous amount of work (seriously)
- Postprocessing and transformation
  - aggregated to two-week averages or sums
  - PCAed some variables to reduce dimensionality down to 3 principal components
  - chose 1-3 fixed lags for each variable according to forecast horizon, data availability
- Data processing challenges: weird data coding, weird data formats, huge data, inconsistent/untimely data updates, real-time processing

Our dataset

- This was an enormous amount of work (seriously)
- Postprocessing and transformation
  - aggregated to two-week averages or sums
  - PCAed some variables to reduce dimensionality down to 3 principal components
  - chose 1-3 fixed lags for each variable according to forecast horizon, data availability
- Data processing challenges: weird data coding, weird data formats, huge data, inconsistent/untimely data updates, real-time processing
- Statistical challenges: spatiotemporal data, correlated predictors, complex dependence structure (careful holdout for cross-validation), non-standard loss function

Our dataset

- This was an enormous amount of work (seriously)
- Postprocessing and transformation
  - aggregated to two-week averages or sums
  - PCAed some variables to reduce dimensionality down to 3 principal components
  - chose 1-3 fixed lags for each variable according to forecast horizon, data availability
- Data processing challenges: weird data coding, weird data formats, huge data, inconsistent/untimely data updates, real-time processing
- Statistical challenges: spatiotemporal data, correlated predictors, complex dependence structure (careful holdout for cross-validation), non-standard loss function
- `SubseasonalRodeo` Dataset: `https://doi.org/10.7910/DVN/IHBANG`

Data matrix

$$
\begin{pmatrix}
\text{lat} & \text{lon} & \text{date} & \text{rhum\_shift30} & \text{pres\_shift30} & \dots & \text{target} \\
47 & 260 & 1979\text{-}02\text{-}09 & 86.539415 & 96061.320731 & \dots & -18.464830 \\
47 & 261 & 1979\text{-}02\text{-}09 & 89.957313 & 96419.183454 & \dots & -18.329887 \\
47 & 262 & 1979\text{-}02\text{-}09 & 92.553695 & 97493.990932 & \dots & -18.289105 \\
48 & 236 & 1979\text{-}02\text{-}09 & 93.731037 & 97277.973493 & \dots & 2.575200 \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots
\end{pmatrix}_{10^6 \times 30}
$$

Our models

- Two regression models

Our models

- Two regression models
  - *MultiLLR* (local linear regression with multitask model selection): uses lagged predictors based on all weather variables, chosen using multitask model selection tailored to the cosine similarity objective
  - *AutoKNN* (k-nearest-neighbors autoregression): uses lagged temperature or precipitation only, and a skill-specific form of nearest neighbor modeling

Our models

- Two regression models
  - *MultiLLR* (local linear regression with multitask model selection): uses lagged predictors based on all weather variables, chosen using multitask model selection tailored to the cosine similarity objective
  - *AutoKNN* (k-nearest-neighbors autoregression): uses lagged temperature or precipitation only, and a skill-specific form of nearest neighbor modeling
- Ensemble of the two models performs better than either model individually

MultiLLR: Local linear regression with multitask model selection

- Subset the training data to dates within 8 weeks of the target date to be predicted

MultiLLR: Local linear regression with multitask model selection

- Subset the training data to dates within 8 weeks of the target date to be predicted
  - This is the “local” part

MultiLLR: Local linear regression with multitask model selection

- Subset the training data to dates within 8 weeks of the target date to be predicted
  - This is the “local” part
- Using backward stepwise selection with linear regression subroutine, choose a common set of relevant predictors for all grid points

MultiLLR: Local linear regression with multitask model selection

- Subset the training data to dates within 8 weeks of the target date to be predicted
  - This is the “local” part
- Using backward stepwise selection with linear regression subroutine, choose a common set of relevant predictors for all grid points
  - This is the “linear regression with multitask model selection” part
  - Don’t expect all features to be relevant at all times of year

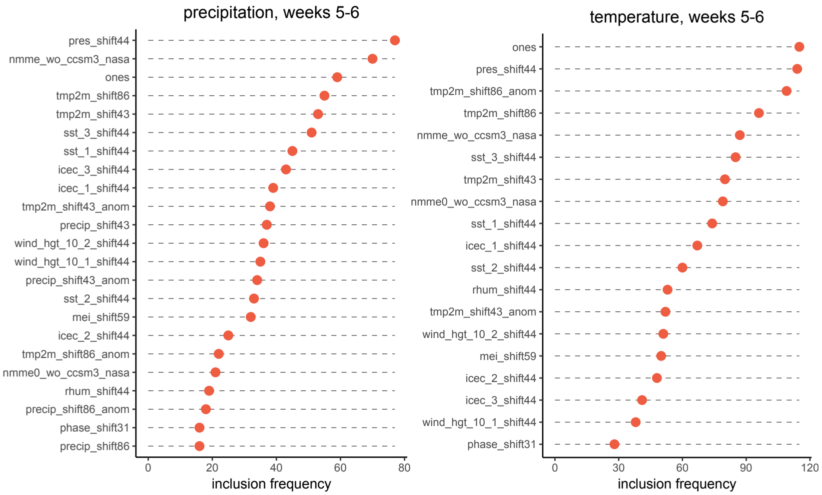
MultiLLR: Local linear regression with multitask model selection

- Subset the training data to dates within 8 weeks of the target date to be predicted
  - This is the “local” part
- Using backward stepwise selection with linear regression subroutine, choose a common set of relevant predictors for all grid points
  - This is the “linear regression with multitask model selection” part
  - Don’t expect all features to be relevant at all times of year
- Backward stepwise has to be customized

MultiLLR: Local linear regression with multitask model selection

- Subset the training data to dates within 8 weeks of the target date to be predicted
  - This is the “local” part
- Using backward stepwise selection with linear regression subroutine, choose a common set of relevant predictors for all grid points
  - This is the “linear regression with multitask model selection” part
  - Don’t expect all features to be relevant at all times of year
- Backward stepwise has to be customized
  - At each step, remove variable that decreases predictive performance the least
  - Predictive performance is the *leave-one-year-out cross-validated cosine similarity* on the target date’s day of year, averaged across all historical years
  - To properly leave one year out around t, need to hold out from 4 weeks before t to 48 weeks after t

Inclusion frequencies of candidate variables



AutoKNN: Multitask k-nearest-neighbor autoregression

- For each target date t, find the 20 most similar historical dates by looking at cosine similarity between anomaly trajectory in the 60 days leading up to t and leading up to each historical date

AutoKNN: Multitask k-nearest-neighbor autoregression

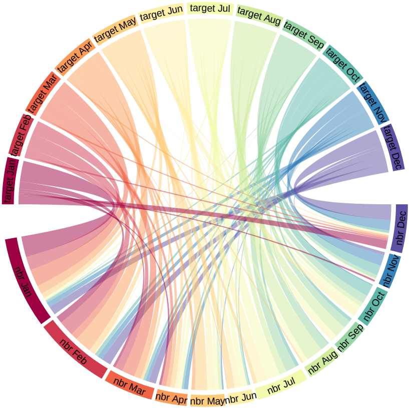
- For each target date t, find the 20 most similar historical dates by looking at cosine similarity between anomaly trajectory in the 60 days leading up to t and leading up to each historical date
- Call the anomalies of the 20 most similar historical dates knn1 through knn20

AutoKNN: Multitask *k*-nearest-neighbor autoregression

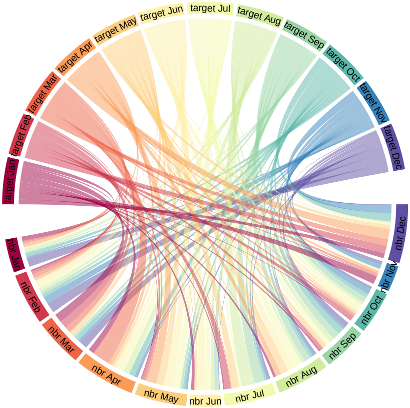
- For each target date t, find the 20 most similar historical dates by looking at cosine similarity between anomaly trajectory in the 60 days leading up to t and leading up to each historical date
- Call the anomalies of the 20 most similar historical dates knn1 through knn20
- Perform weighted local linear regression using knn1 through knn20 and fixed lagged measurements of temperature or precipitation to predict future anomaly

Learned nearest neighbors

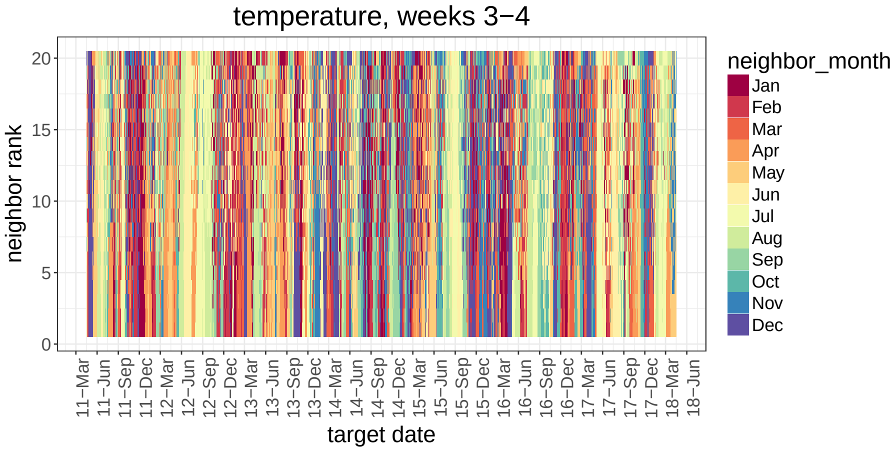
Precipitation



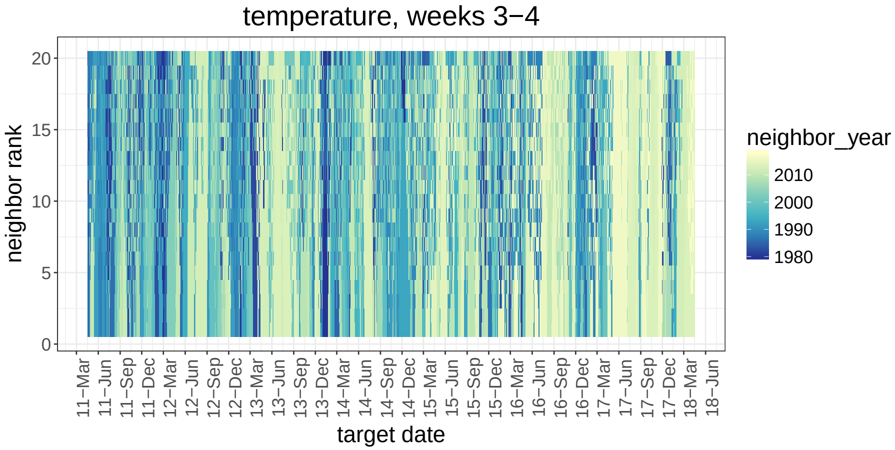
Temperature



Learned nearest neighbors



Learned nearest neighbors



Ensemble of the two models

- We ensemble by averaging the ℓ_2-normalized forecasted anomalies:

$$\hat{\mathbf{a}}_{\text{ensemble}} = \frac{1}{2}\frac{\hat{\mathbf{a}}_{\text{LLR}}}{\|\hat{\mathbf{a}}_{\text{LLR}}\|_2} + \frac{1}{2}\frac{\hat{\mathbf{a}}_{\text{KNN}}}{\|\hat{\mathbf{a}}_{\text{KNN}}\|_2}$$

Ensemble of the two models

- We ensemble by averaging the ℓ_2-normalized forecasted anomalies:

$$\hat{\mathbf{a}}_{\text{ensemble}} = \frac{1}{2}\frac{\hat{\mathbf{a}}_{\text{LLR}}}{\|\hat{\mathbf{a}}_{\text{LLR}}\|_2} + \frac{1}{2}\frac{\hat{\mathbf{a}}_{\text{KNN}}}{\|\hat{\mathbf{a}}_{\text{KNN}}\|_2}$$

- This is guaranteed to improve the average skill, thanks to. . .

Ensemble of the two models

- We ensemble by averaging the ℓ_2-normalized forecasted anomalies:

$$\hat{\mathbf{a}}_{\text{ensemble}} = \frac{1}{2}\frac{\hat{\mathbf{a}}_{\text{LLR}}}{\|\hat{\mathbf{a}}_{\text{LLR}}\|_2} + \frac{1}{2}\frac{\hat{\mathbf{a}}_{\text{KNN}}}{\|\hat{\mathbf{a}}_{\text{KNN}}\|_2}$$

- This is guaranteed to improve the average skill, thanks to. . .

Proposition

Consider an observed anomaly vector $\mathbf{a}$ and m distinct forecast anomaly vectors $(\hat{\mathbf{a}}_i)_{i=1}^m$. For any vector of weights $\mathbf{p} \in \mathbb{R}^m$ with $\sum_{i=1}^m p_i = 1$ and $p_i \geq 0$, let

$$\bar{\mathbf{a}}_{(\mathbf{p})} = \sum_{i=1}^{m} p_i \frac{\hat{\mathbf{a}}_i}{\|\hat{\mathbf{a}}_i\|}$$

be the weighted average of the ℓ_2-normalized forecast anomalies. Then,

$$\left| \sum_{i=1}^{m} p_i \cos(\hat{\mathbf{a}}_i, \mathbf{a}) \right| \leq |\cos(\bar{\mathbf{a}}_{(\mathbf{p})}, \mathbf{a})|.$$

Results

- In the contest period (2017-2018), our models beat both of the contest baselines (and the top competitor) by a lot
  - unfortunately, only arrived at these model late in the competition

Results

- In the contest period (2017-2018), our models beat both of the contest baselines (and the top competitor) by a lot
  - unfortunately, only arrived at these model late in the competition
- In a historical evaluation period (2011-2017), our models beat a reconstructed baseline by a lot

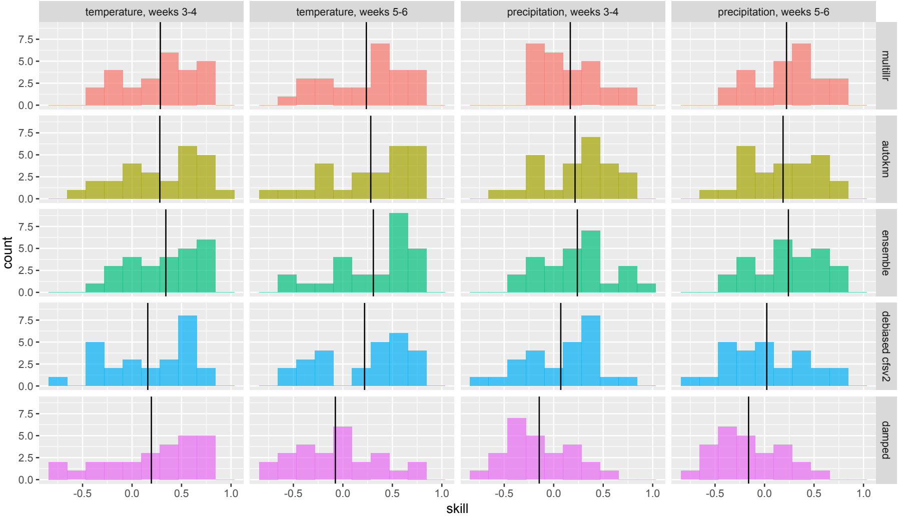
Results

- In the contest period (2017-2018), our models beat both of the contest baselines (and the top competitor) by a lot
  - unfortunately, only arrived at these model late in the competition
- In a historical evaluation period (2011-2017), our models beat a reconstructed baseline by a lot
- Ensembling the two models helps significantly

Contest period, 2017-2018

task	LLR	KNN	ensemble	cfsv2	damped	top competitor
temp, weeks 3-4	0.2856	0.2807	**0.3414**	0.1589	0.1952	0.2855
temp, weeks 5-6	0.2371	0.2817	**0.3077**	0.2192	-0.0762	0.2357
precip, weeks 3-4	0.1675	0.2156	**0.2388**	0.0713	-0.1463	0.2144
precip, weeks 5-6	0.2219	0.1870	**0.2412**	0.0227	-0.1613	0.2162

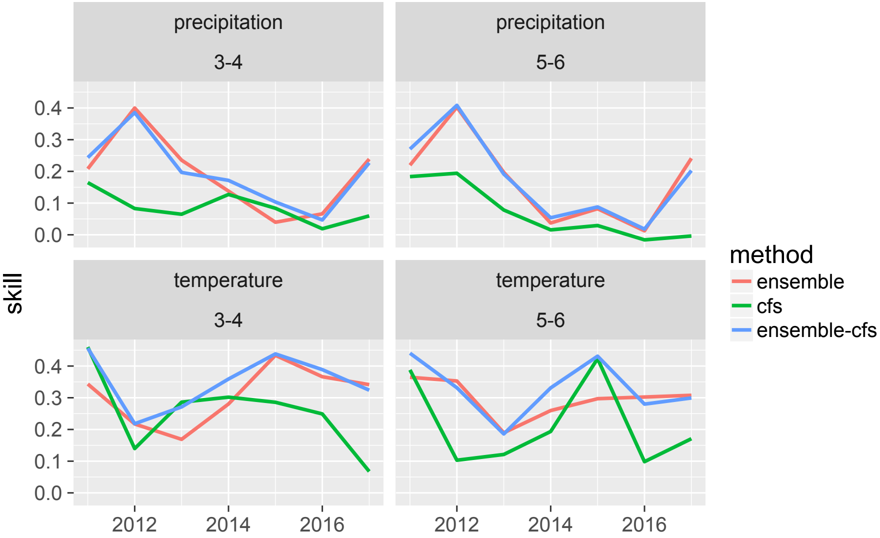
Contest period, 2017-2018



Historical evaluation period, 2011-2017

task	LLR	KNN	ensemble	cfsv2	ens-cfsv2
temp, weeks 3-4	0.2230	0.3111	0.3073	0.2557	**0.3508**
temp, weeks 5-6	0.2204	0.2810	0.2962	0.2142	**0.3279**
precip, weeks 3-4	0.1573	0.1513	0.1893	0.0860	**0.1964**
precip, weeks 5-6	0.1312	0.1403	0.1703	0.0691	**0.1755**

Historical evaluation period, 2011-2017



Conclusion

- Subseasonal Forecasting is an important and largely unsolved problem in weather prediction

Conclusion

- Subseasonal Forecasting is an important and largely unsolved problem in weather prediction
- Simple statistical models can significantly improve subseasonal forecasting compared to physics-based models

Conclusion

- Subseasonal Forecasting is an important and largely unsolved problem in weather prediction
- Simple statistical models can significantly improve subseasonal forecasting compared to physics-based models
- Ensembling statistical and physics-based forecasts produce further improvements

Conclusion

- Subseasonal Forecasting is an important and largely unsolved problem in weather prediction
- Simple statistical models can significantly improve subseasonal forecasting compared to physics-based models
- Ensembling statistical and physics-based forecasts produce further improvements
- We've released the dataset on Dataverse (`https://doi.org/10.7910/DVN/IHBANG`) and our paper on arXiv (`https://arxiv.org/abs/1809.07394`)

Conclusion

- Subseasonal Forecasting is an important and largely unsolved problem in weather prediction
- Simple statistical models can significantly improve subseasonal forecasting compared to physics-based models
- Ensembling statistical and physics-based forecasts produce further improvements
- We've released the dataset on Dataverse (`https://doi.org/10.7910/DVN/IHBANG`) and our paper on arXiv (`https://arxiv.org/abs/1809.07394`)
- More sophisticated modeling approaches can almost certainly do even better. Try your own!