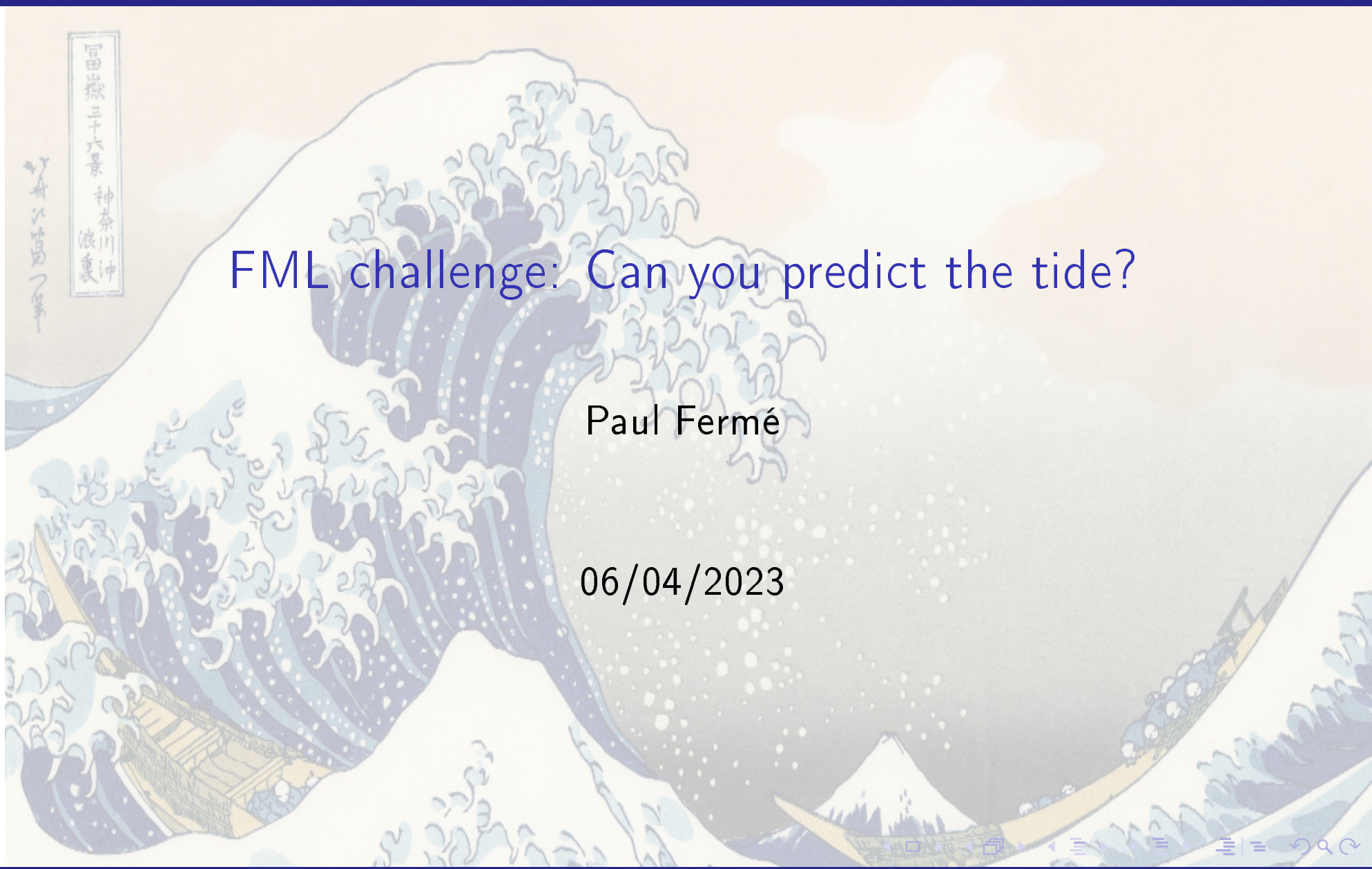


# FML challenge: Can you predict the tide?

Paul Fermé

06/04/2023

# Predict the sea *surge*

- Variations of sea level composed of *tide* and *surge*.

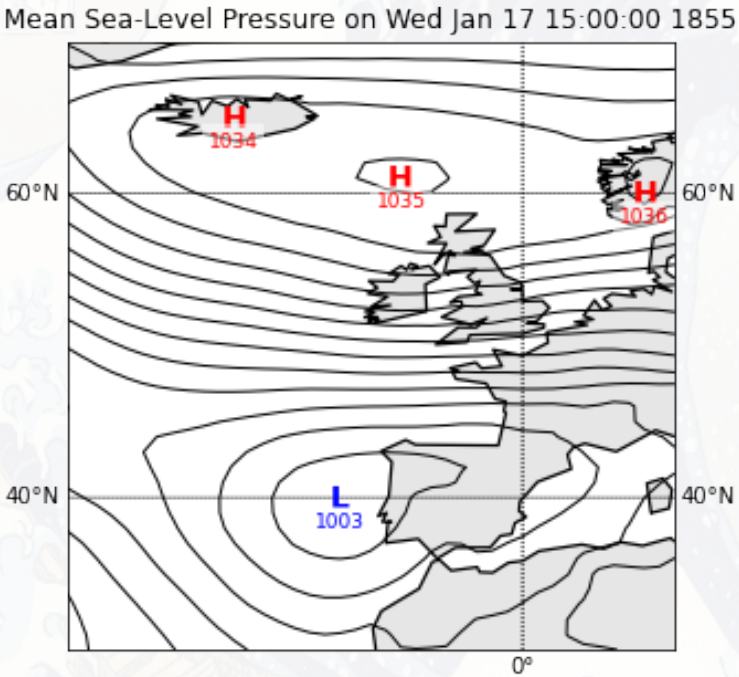
# Predict the sea *surge*

- Variations of sea level composed of *tide* and *surge*.
- *Tide* composed of astronomical tide and radiational tide. *Easily predicted.*

# Predict the sea *surge*

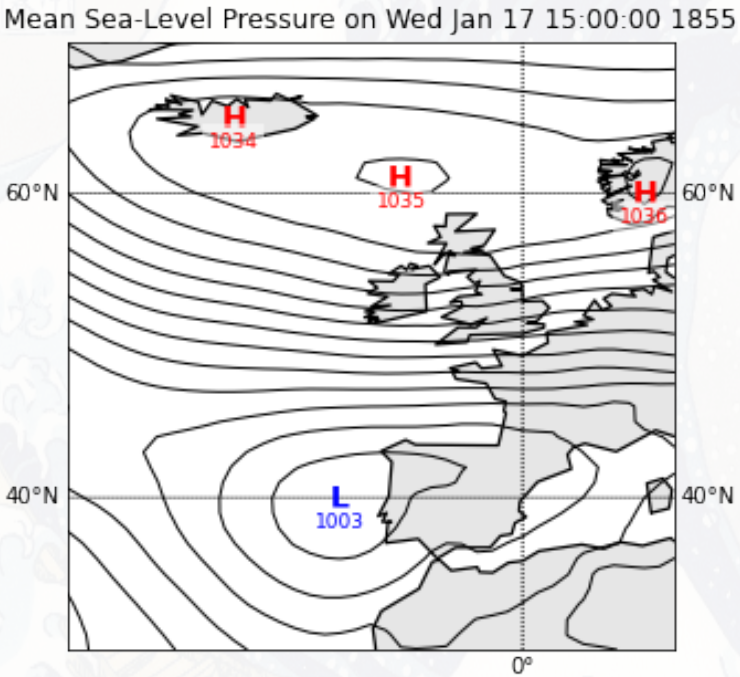
- Variations of sea level composed of *tide* and *surge*.
- *Tide* composed of astronomical tide and radiational tide. *Easily predicted.*
- *Surge* is the difference between observed sea level and *tide*. *Critical importance for safety reasons.*

# Data description




- Sea-level pressure field (SLP).
- Sea surges of two (unknown) cities.

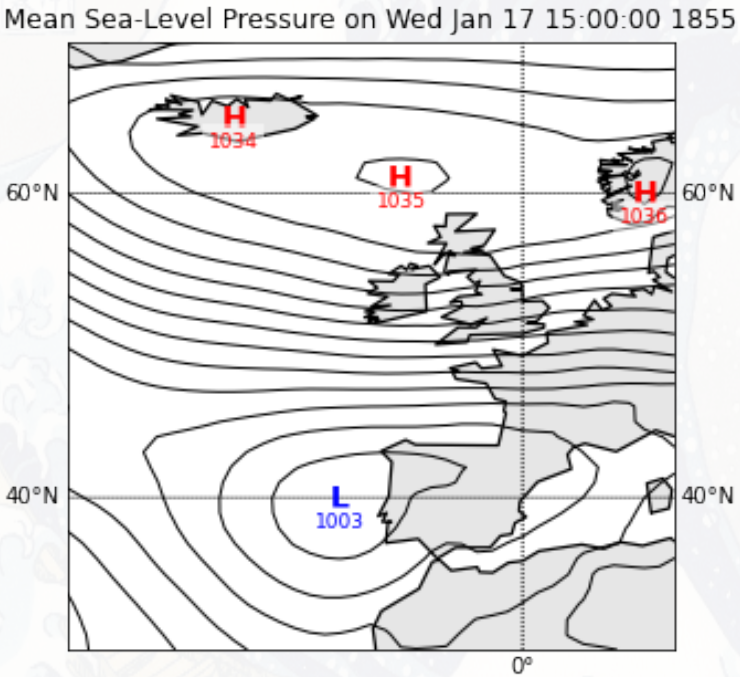
# Data description




- Sea-level pressure field (SLP).
- Sea surges of two (unknown) cities.

- Data from 1950s to 2010s.
- Windows of five days.
- SLP: $41 \times 41$ grid, every 3 hours.
- Surges: Normalized, every 12 hours.

# Data description



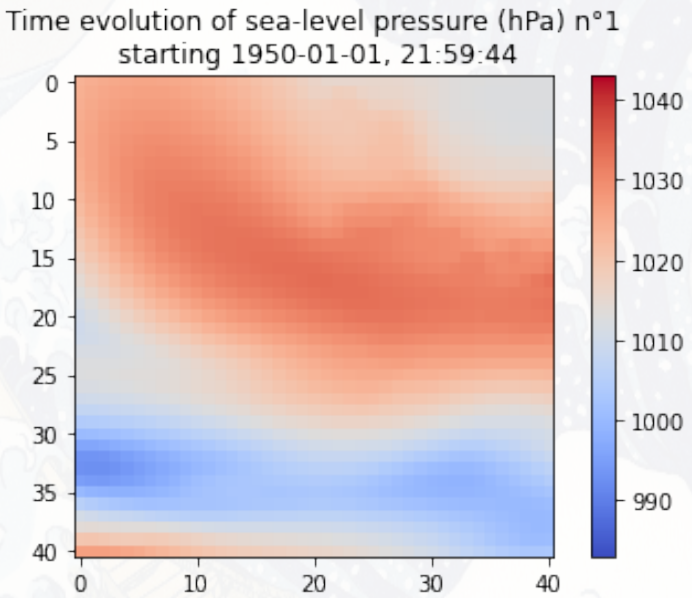

- Sea-level pressure field (SLP).
- Sea surges of two (unknown) cities.

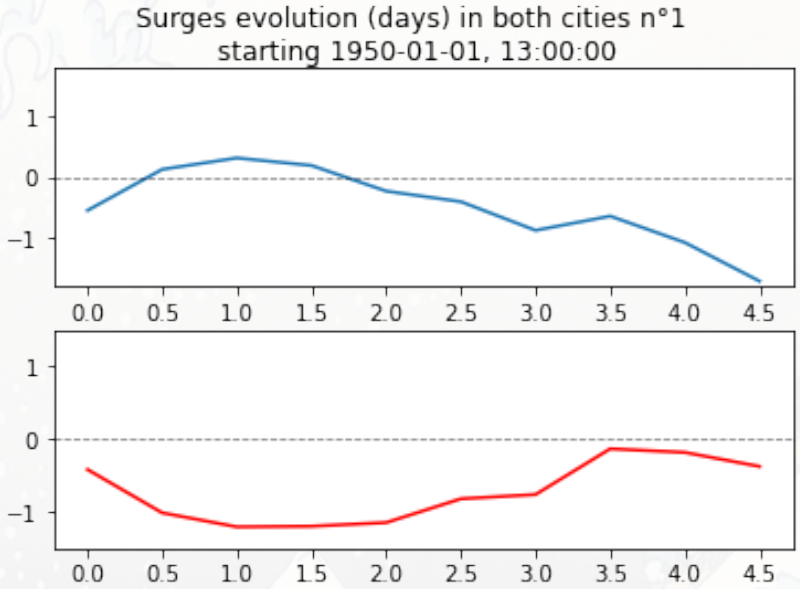
- Data from 1950s to 2010s.
- Windows of five days.
- SLP: $41 \times 41$ grid, every 3 hours.
- Surges: Normalized, every 12 hours.

- Output: Surges on next five days.

# Example data

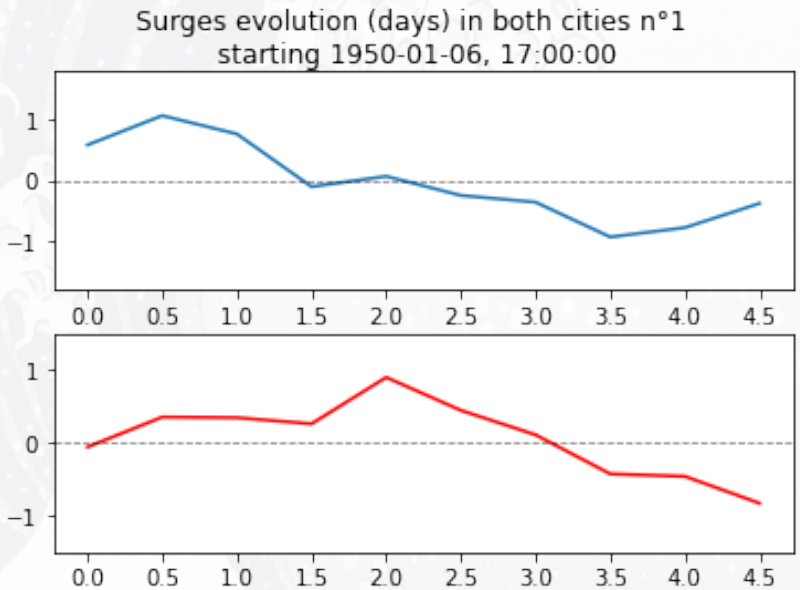
Input:



# Example data

Output:



# Benchmark: K-nearest neighbors

- Metric: $L^2$ loss function, weight $(1, 0.9, \ldots, 0.1)$.

# Benchmark: K-nearest neighbors

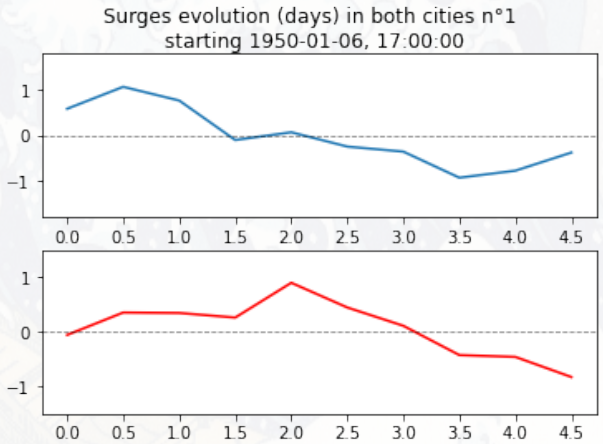
- Metric: $L^2$ loss function, weight $(1, 0.9, \ldots, 0.1)$.
- Benchmark: K-nearest neighbors on SLP at time $t$ and $t - 24h$, with $K = 40$, ignoring other SLPs, time, input surges.

# Benchmark: K-nearest neighbors

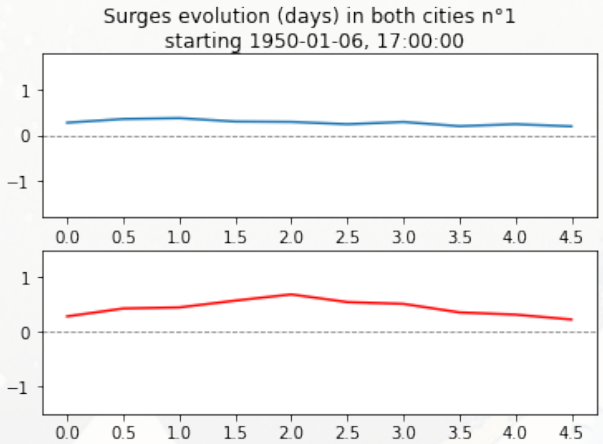
- Metric: $L^2$ loss function, weight $(1, 0.9, \ldots, 0.1)$.
- Benchmark: K-nearest neighbors on SLP at time $t$ and $t - 24h$, with $K = 40$, ignoring other SLPs, time, input surges.
- Achieves a score of 0.7720, ie. 23% of variance explained.

# Benchmark: K-nearest neighbors

- Metric: $L^2$ loss function, weight $(1, 0.9, \ldots, 0.1)$.
- Benchmark: K-nearest neighbors on SLP at time $t$ and $t - 24h$, with $K = 40$, ignoring other SLPs, time, input surges.
- Achieves a score of 0.7720, ie. 23% of variance explained.




True surges




Benchmark prediction

# Our method

- Normalize the data:
  1. Normalize SLPs and time (input surges already normalized).
  2. Transform input surges to implement the weighted metric with a standard $L^2$ norm.

# Our method

- Normalize the data:
  1. Normalize SLPs and time (input surges already normalized).
  2. Transform input surges to implement the weighted metric with a standard $L^2$ norm.
- Dimension reduction on SLPs with PCA.

# Our method

- Normalize the data:
  1. Normalize SLPs and time (input surges already normalized).
  2. Transform input surges to implement the weighted metric with a standard $L^2$ norm.
- Dimension reduction on SLPs with PCA.
- Lasso regression for city 1 (resp. 2) on:
  1. Reduced SLPs.
  2. Input surges city 1 (resp. 2).
  3. Time.

# Our method

- Normalize the data:
  1. Normalize SLPs and time (input surges already normalized).
  2. Transform input surges to implement the weighted metric with a standard $L^2$ norm.
- Dimension reduction on SLPs with PCA.
- Lasso regression for city 1 (resp. 2) on:
  1. Reduced SLPs.
  2. Input surges city 1 (resp. 2).
  3. Time.
- Transform back output surges.

# Our method

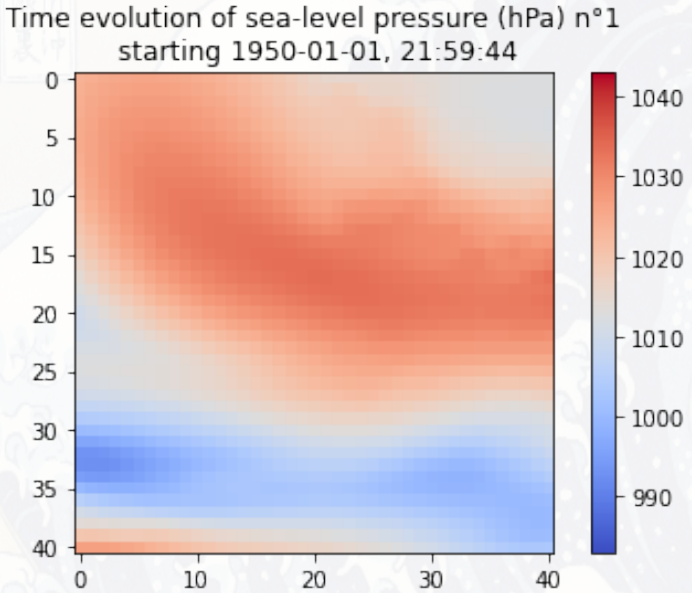
- Normalize the data:
  1. Normalize SLPs and time (input surges already normalized).
  2. Transform input surges to implement the weighted metric with a standard $L^2$ norm.
- Dimension reduction on SLPs with PCA.
- Lasso regression for city 1 (resp. 2) on:
  1. Reduced SLPs.
  2. Input surges city 1 (resp. 2).
  3. Time.
- Transform back output surges.
- With the right hyperparameters, achieves a score of 0.4993, ie. more than 50% of the variance explained!

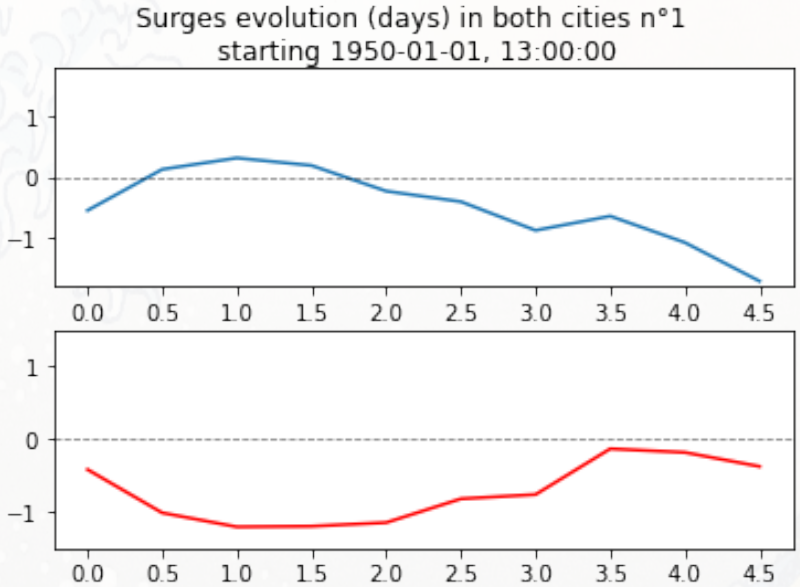
# Our method

- Normalize the data:
  1. Normalize SLPs and time (input surges already normalized).
  2. Transform input surges to implement the weighted metric with a standard $L^2$ norm.
- Dimension reduction on SLPs with PCA.
- Lasso regression for city 1 (resp. 2) on:
  1. Reduced SLPs.
  2. Input surges city 1 (resp. 2).
  3. Time.
- Transform back output surges.

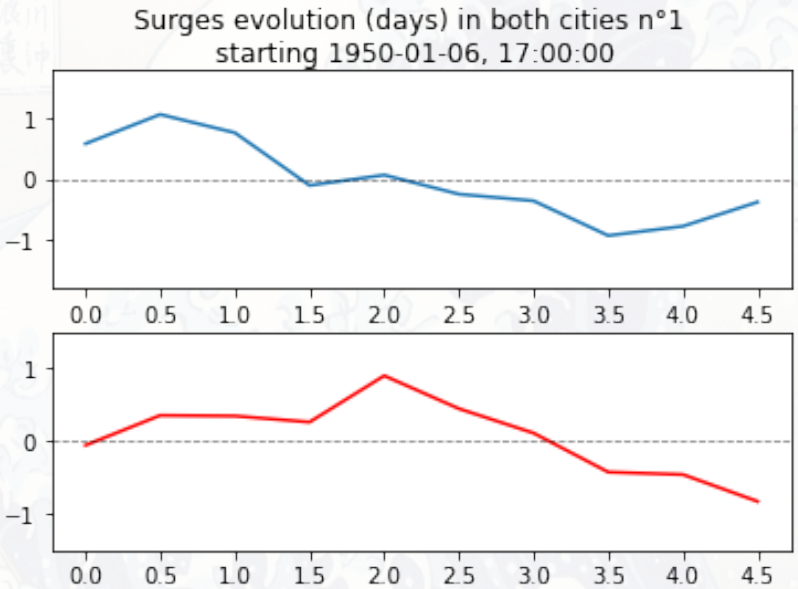
- With the right hyperparameters, achieves a score of 0.4993, ie. more than 50% of the variance explained!
- $<$ 6% difference with the best solution online (0.4419).

# On an example

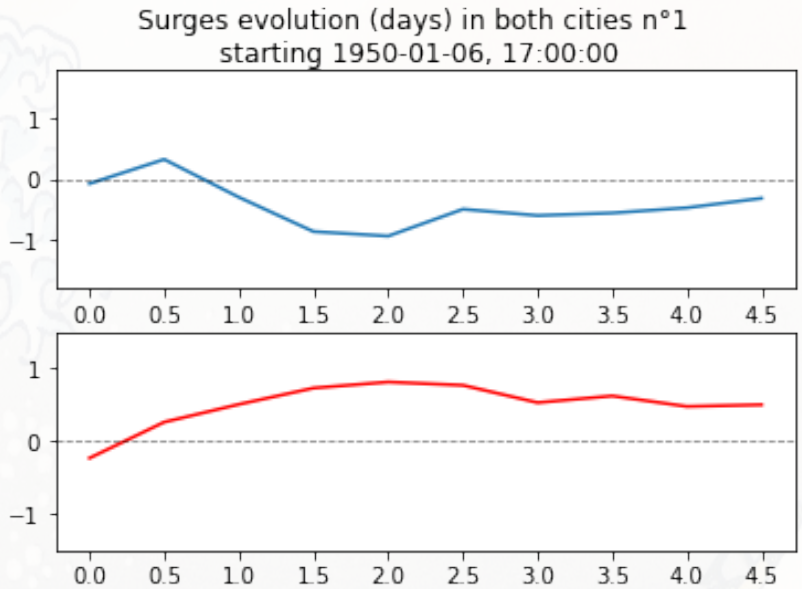


# On an example

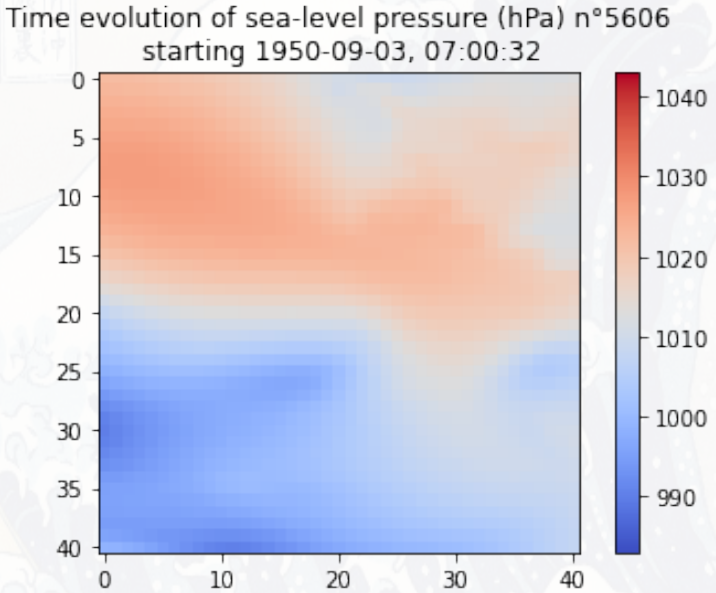


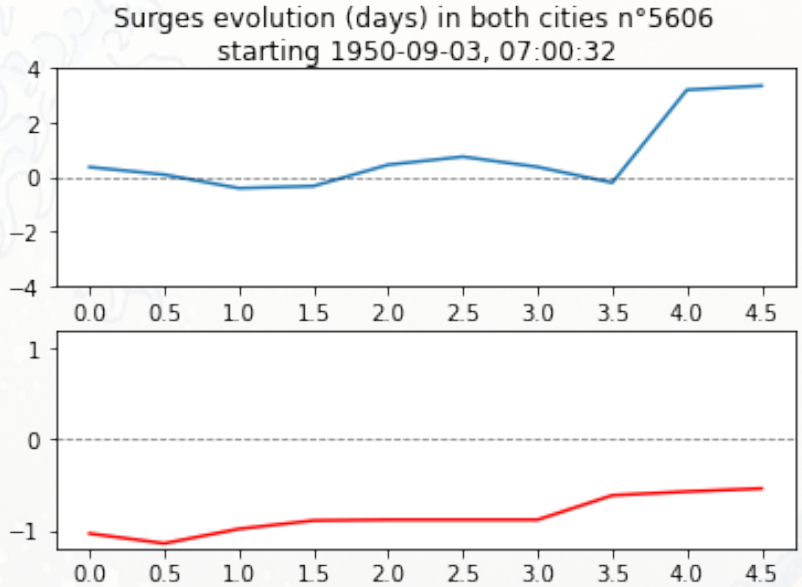

True surges



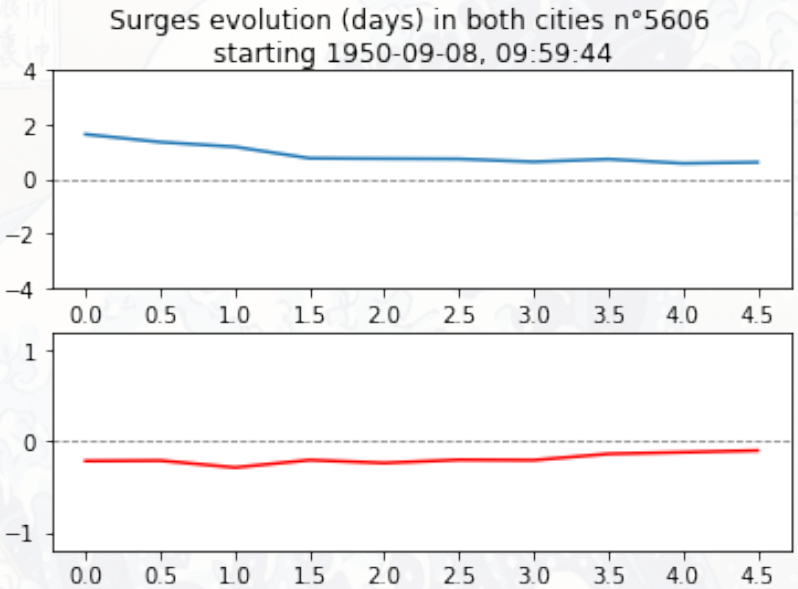

Our prediction

# Comparison to benchmark

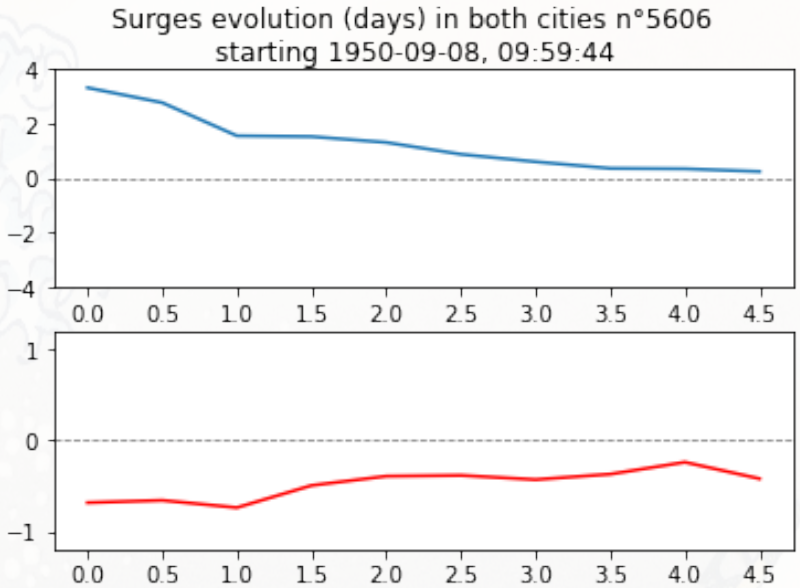


# Comparison to benchmark




Benchmark




Our prediction

# Model interpretation

Since linear regression model, we can interpret coefficients:

# Model interpretation

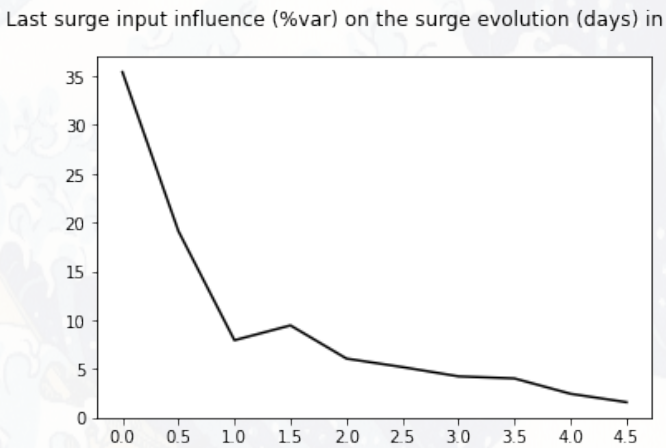
Since linear regression model, we can interpret coefficients:

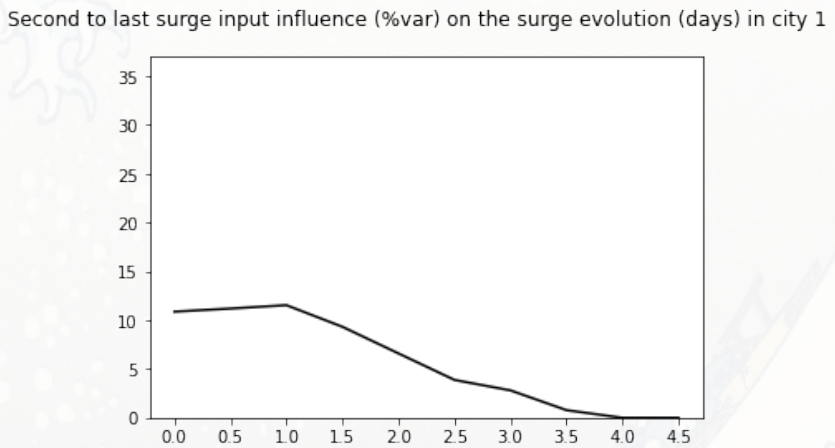
- SLPs:
  1. Hard to interpret: work on (198) features extracted with PCA.
  2. Mean 0.25%, standard deviation 0.29%.

# Model interpretation

Since linear regression model, we can interpret coefficients:

- SLPs:
  1. Hard to interpret: work on (198) features extracted with PCA.
  2. Mean 0.25%, standard deviation 0.29%.
- Input surges:

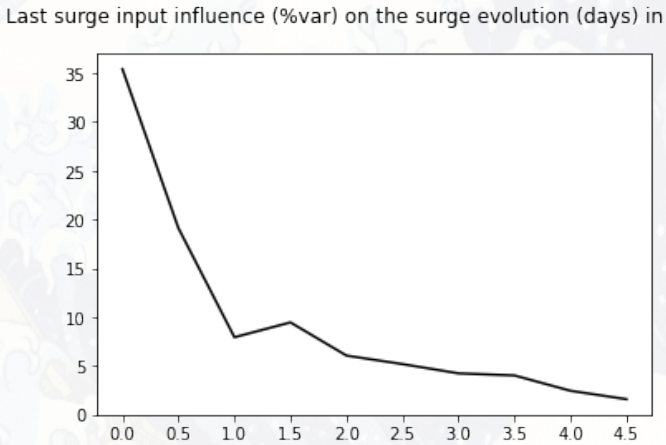


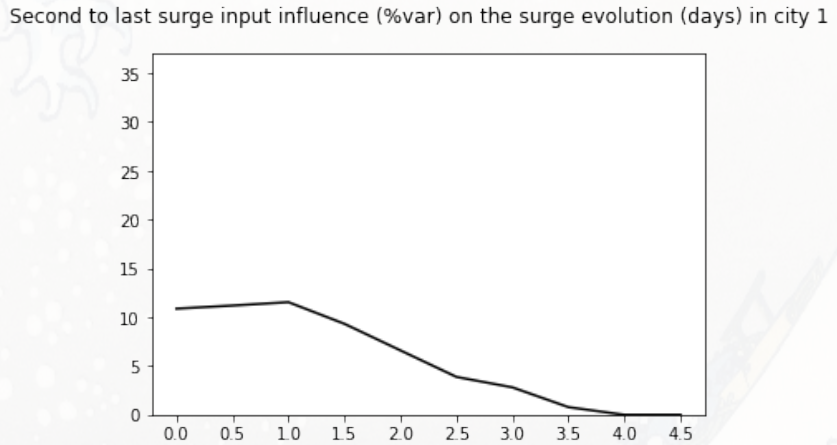



# Model interpretation

Since linear regression model, we can interpret coefficients:

- SLPs:
  1. Hard to interpret: work on (198) features extracted with PCA.
  2. Mean 0.25%, standard deviation 0.29%.
- Input surges:




- Time: 6.6% (resp. 5.3%) variance increase every 10 years in city 1 (resp. 2). *Global warming?*

# Issues and other attempted methods

- Data correlation:
  1. Windows of data over 5 days, but samples every day.
  2. If one takes only unoverlapped data, test efficiency decreases.
  3. Taking subset of 40 time steps of SLPs does not work either.

# Issues and other attempted methods

- Data correlation:
  1. Windows of data over 5 days, but samples every day.
  2. If one takes only unoverlapped data, test efficiency decreases.
  3. Taking subset of 40 time steps of SLPs does not work either.

- Improve KNN?
  1. Tweaking hyperparameters: no real increase over benchmark.
  2. Wasserstein distance (optimal transport): too hard to compute on whole SLPs.
  3. Wind (gradient) computation seems to add only noise.

# Issues and other attempted methods

- Data correlation:
  1. Windows of data over 5 days, but samples every day.
  2. If one takes only unoverlapped data, test efficiency decreases.
  3. Taking subset of 40 time steps of SLPs does not work either.

- Improve KNN?
  1. Tweaking hyperparameters: no real increase over benchmark.
  2. Wasserstein distance (optimal transport): too hard to compute on whole SLPs.
  3. Wind (gradient) computation seems to add only noise.

- Neural Network instead of Lasso regression? Either overfits or behaves poorly compared to Lasso.

# Conclusion and further directions

- We have developped a simple method, using only PCA and Lasso regression, to predict sea surges.
- Explains $> 50\%$ of variance (benchmark 23%, best 56%).
- Model predicts 6% increase/10 years: *climate change*.

# Conclusion and further directions

- We have developped a simple method, using only PCA and Lasso regression, to predict sea surges.
- Explains $> 50\%$ of variance (benchmark 23%, best 56%).
- Model predicts 6% increase/10 years: *climate change*.

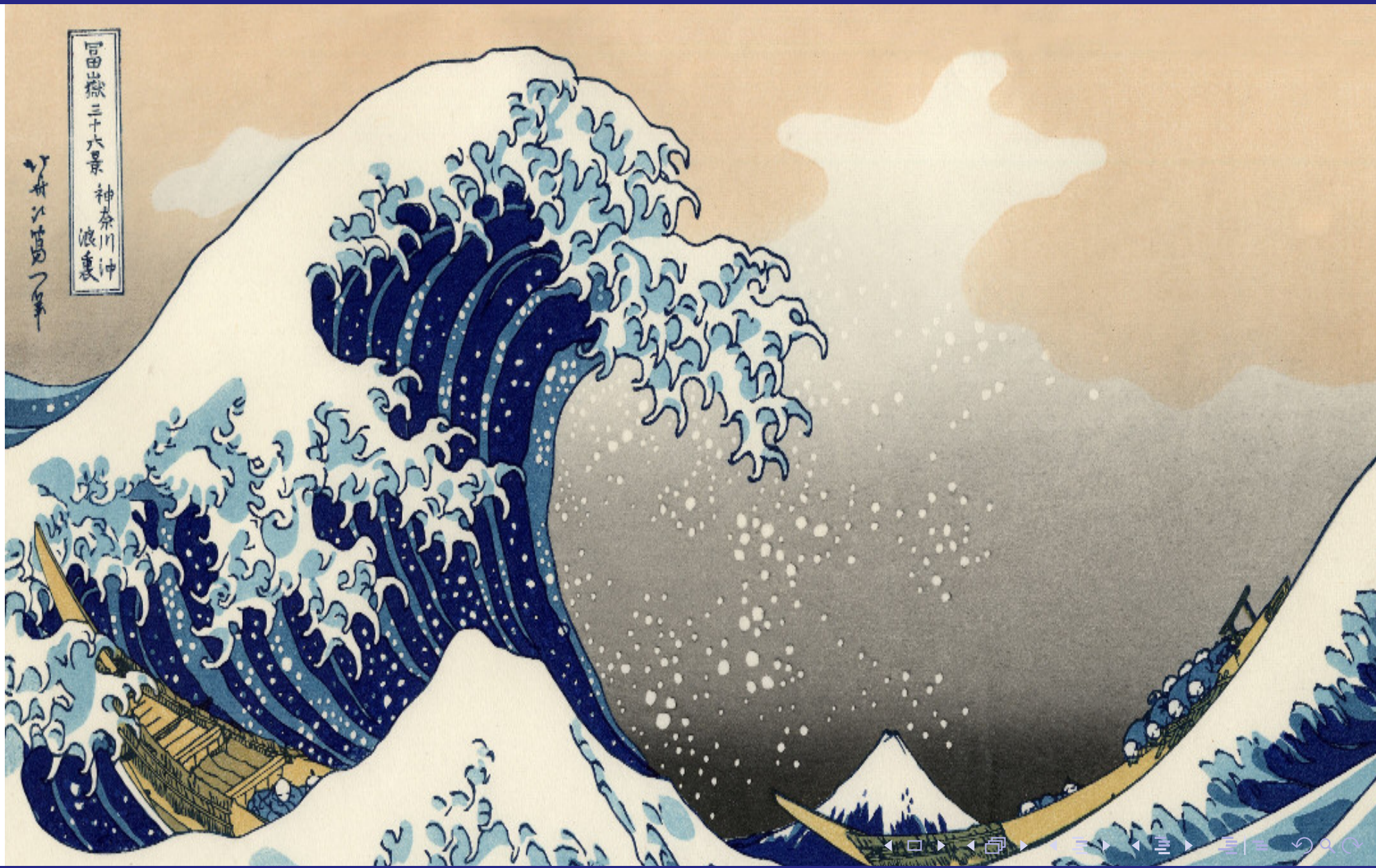
- Preprocessing of SLPs in another (and more significant) way than PCA?
- Time series consideration?

# Conclusion and further directions

- We have developped a simple method, using only PCA and Lasso regression, to predict sea surges.
- Explains $> 50\%$ of variance (benchmark 23%, best 56%).
- Model predicts 6% increase/10 years: *climate change*.

- Preprocessing of SLPs in another (and more significant) way than PCA?
- Time series consideration?

- *Thank you!*


富嶽三十六景
神奈川沖
浪裏