Spatial Echo State Networks for Oceanographic Data Analysis

Master’s Thesis by Hannah Melzer
Supervised by Prof. Stephan Juricke,
Prof. Markus Jochum & Prof. James Avery

Overview

GEOMAR

- Motivation
- Machine Learning Models → Reservoir Computing
- A spatial loss function
- Defining normality → Anomaly detection
- Data & Test region
- Results I - Modelling the Kuroshio
- Results II - Anomaly Detection
- Future Work & Ideas

Motivation - Exploring the Unexplored

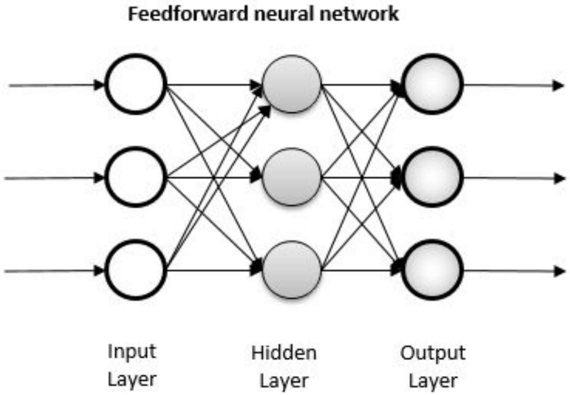
"The virtual oceans are just as unexplored as the real ones."
- Markus Jochum

- modern ocean models produce **several TB** of data
- with sometimes **40+ variables**
- over **100-1000s of years**

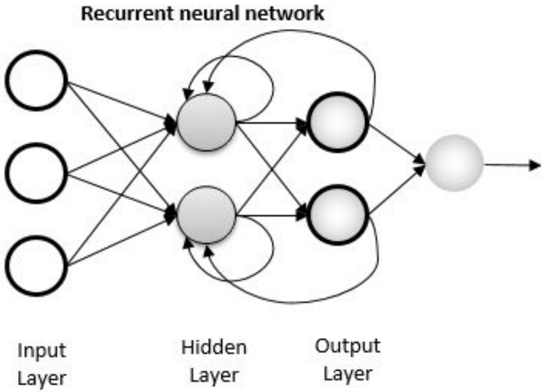
→ analysing this will take an enormous amount of time/ or a lot of students

→ need for an **automated anomaly detection**

Machine Learning models



want to remember past input



Problems: long training times, vanishing gradients …

Solution

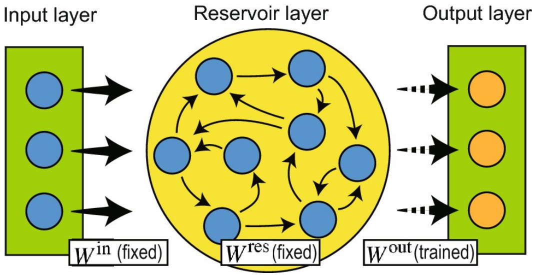
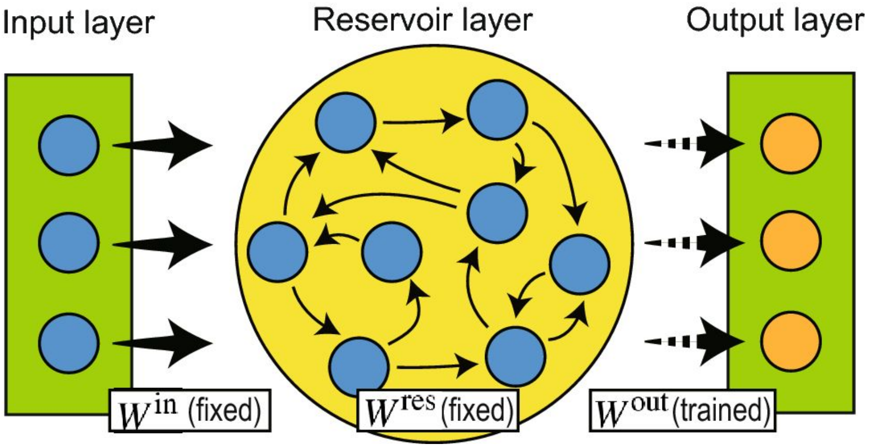


image sources: Researchgate, Nature

Reservoir Computing - Echo State Network



- weights in reservoir are **FIXED**
- only output layer is trained

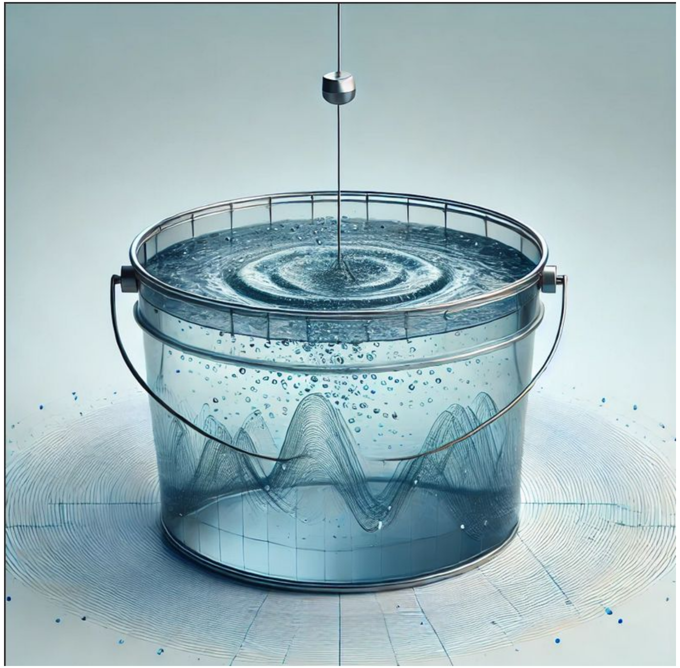
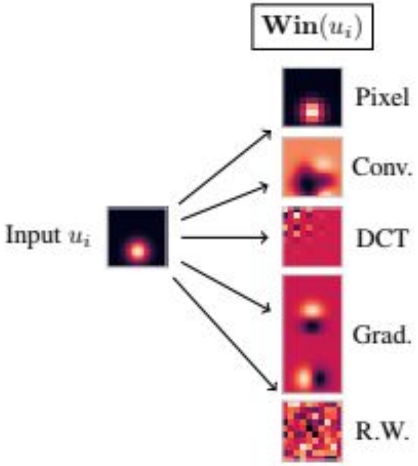
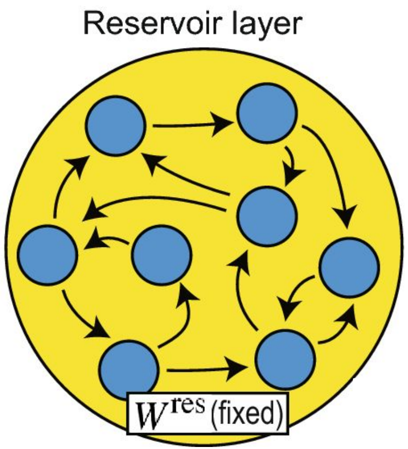
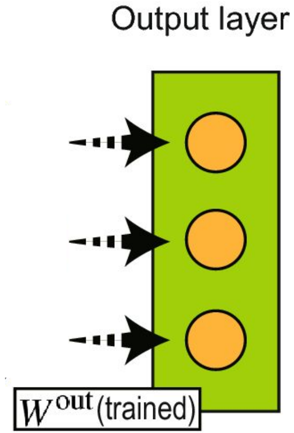


image sources: Nature, ChatGPT

Tuning of the ESN







- different input mappings

- spectral radius (largest absolute eigenvalue)
- select size, connections etc.

- least square fit

image sources: Nature, Heim and Avery, 2019

A spatial loss function - IMED (IMage Euclidean Distance metric)

GEOMAR

- ML methods apply **optimization to minimize** their mistakes i.e. **loss**
  → one of the most crucial choices
- Euclidean distance compares pixel to pixel
  → even **slight displacements** lead to seemingly **huge errors**
  → **spatially sensitive** metric to account for **neighbourhood structure**
  → smoothing with **Gaussian blur**
- Wang et al. 2005

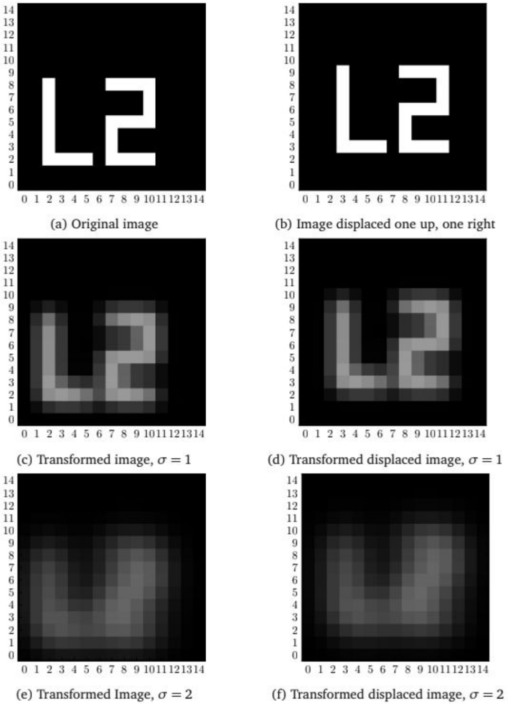


image source: Jacob Ungar Felding

HELMHOLTZ

Defining Normality

- Assumption: **prediction** = expected i.e. "**normal**" behaviour
  → identify **normality** with **predictability**
- **two** moving averages with **different window sizes**

$$\underbrace{\tau_n}_{\substack{\textbf{small}\text{ window that} \\ \textbf{conserves errors}}} << \underbrace{\tau_m}_{\substack{\textbf{large}\text{ window that} \\ \textbf{smoothes out errors}}}$$

→ error relative to recent history i.e. possible **anomaly**

Data & Test Region

- 1/10 deg CESM data
- 17 years 5 daily
- Kuroshio
  - typical Large Meander (tLM)
  - offshore non-large meander (oNLM)
  - nearshore non-large meander (nNLM)

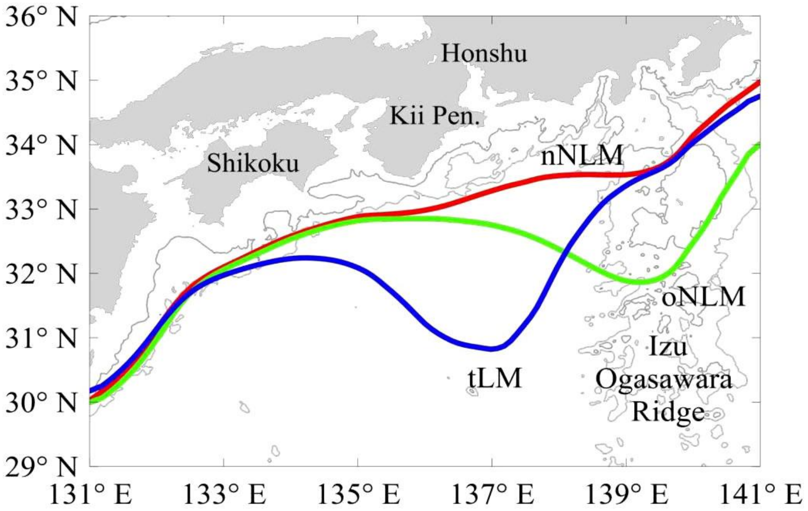
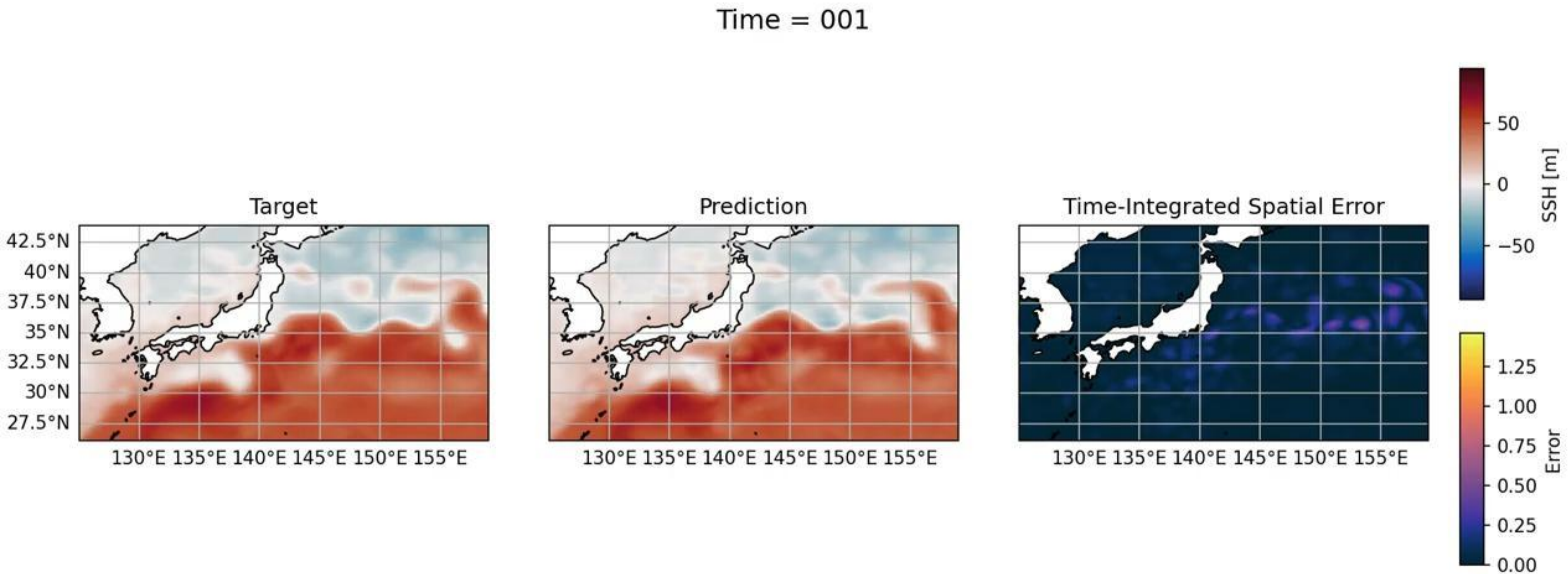
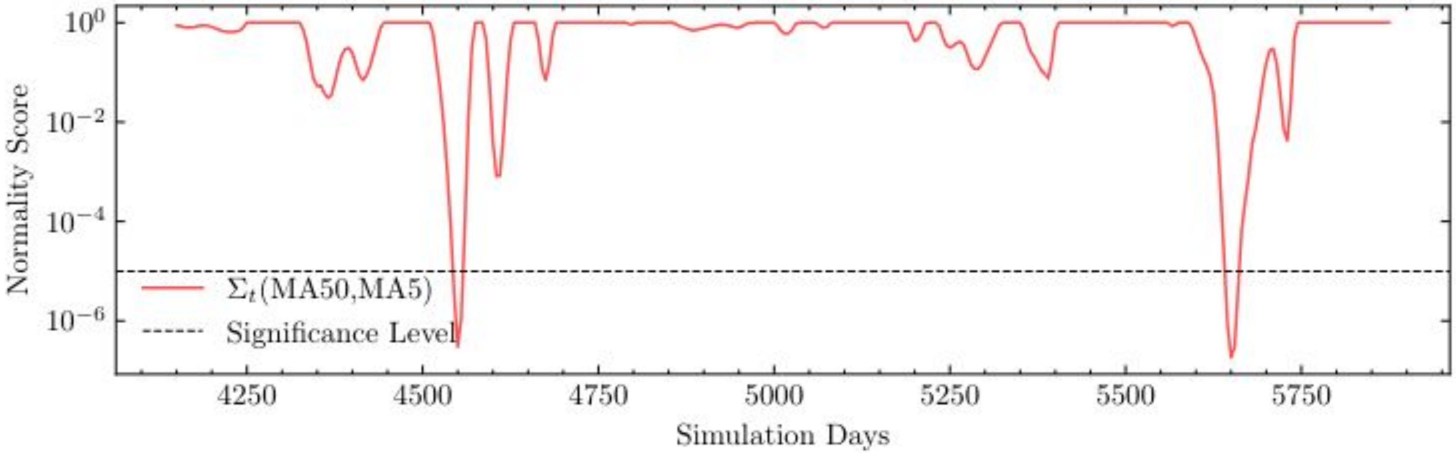


image source: Kawabe 1995

Results I - Modelling the Kuroshio



Results II - Detecting Anomalies



Future Work & Ideas (?)

- Extending to other regions e.g. Agulhas Leakage, Gulf stream, Malvinas Confluence
- applying the model to other properties like density
- global multi scale analysis

Thank you for your attention!

Questions?

Anomaly Score (after Ahmad et al. 2017)

$$A(t) = 1 - Q\left(\frac{\mu_{\tau_1}(t) - \mu_{\tau_2}(t)}{\sigma_{\tau_1}(t)}\right)$$

Here, the Q-function is the Gaussian tail distribution function and can be expressed using the Gaussian error function erf(·):

$$Q(x) = \frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) = \frac{1}{2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right)$$

credit: Jacob Ungar Felding

Overview

- Motivation
-