Probabilistic forecasting of heat waves with deep learning

G. Miloshevich

1Departement de Physique
Ecole Normale Superieure de Lyon

Institut Rhônalpin des systèmes complexes

Machine Learning and sampling methods for climate and physics, 2022
Machine Learning (ML) for extreme events

- The regional impact of climate change remains to be explored[1]
- Extreme events, like heat waves, important but rare
- Forecasting with Artificial Neural Networks (ANNs)[2][3]

1 Intro to Machine Learning (ML)
Outline

1 Intro to Machine Learning (ML)

2 ML in computational Earth sciences
Outline

1. Intro to Machine Learning (ML)
2. ML in computational Earth sciences
3. Predicting Heat Waves (HW) with Deep Learning (DL)
1 Intro to Machine Learning (ML)

2 ML in computational Earth sciences

3 Predicting Heat Waves (HW) with Deep Learning (DL)

4 Future work and conclusions
Outline

1. Intro to Machine Learning (ML)

2. ML in computational Earth sciences

3. Predicting Heat Waves (HW) with Deep Learning (DL)

4. Future work and conclusions
ANNs: image, speech recognition, games

- ML consists of various fields: [4]
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

ANNs: image, speech recognition, games

- ML consists of various fields: [4]
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Components of ANNs:
  - Hyperparameters $\theta$, e.g. weights $w_i$
  - Nonlinear activation function
  - Loss function $E(\theta) = C(X, g(\theta))$
  - Backpropagation to minimize loss [5]

$$\theta_{t+1} = \theta_t - \eta_t \nabla \theta \sum_{i \in B_k} e_i (X_i, \theta) \quad (1)$$

- Universal function approximators [6]

---


Figure: architecture
Outline

1. Intro to Machine Learning (ML)
2. ML in computational Earth sciences
3. Predicting Heat Waves (HW) with Deep Learning (DL)
4. Future work and conclusions
From pattern recognition to physical models

- Early work of Bjerknes to the method of analogues Lorenz[7]
- Success of physical models over pattern recognition, 1950s onwards
- The end of Dennard scaling: arithmetic speed levels off

Figure: Analogue method

From physical models to pattern recognition

- Success of ML in long-term prediction such as ENSO [8]
- Will ML replace or morph with physical modeling? [9]

Figure: Nino3.4 indexes for an 18-month-lead

Studying extremes with models vs ML

- **General Circulation Models (GCMs)** when used for extremes of: \[10\]
  - at the regional scale, are still limited by the rarity of events
  - For uncertainty quantification larger multi-model ensembles wanted

**Figure: European heat wave 2003**

**Figure: Changes in temperatures**\[11\]

\[10\] S. Seneviratne et al., A Special Report of Working Groups I and II of the IPCC (2012)

\[11\] S. E. Perkins, Atmospheric Research (2015)
Outline

1. Intro to Machine Learning (ML)
2. ML in computational Earth sciences
3. Predicting Heat Waves (HW) with Deep Learning (DL)
4. Future work and conclusions
Scandinavian blocking: HW onset

- Rossby wave breaking and blocking
- Advection: persistent anticyclonic anomaly
  \[ V = \frac{k}{f} \times \nabla z \]  
  \[ z(p) = R \int_p^{p_s} \frac{T}{g} \frac{dp}{p} \]  
- Dry soil contributes to heating due to lack of latent heat

Figure: Scandinavia: Average temperature
Figure: Temperature, geopotential (ECMWF)
Summer HWs over France: definition

- HW: extreme of space-time averaged temperature anomalies:

\[ A_T(t) = \frac{1}{T} \int_t^{t+T} \frac{1}{|D|} \int_D (T_{2m} - \mathbb{E}(T_{2m})) (\vec{r}, u) \, d\vec{r} \, du \]  

\[ (4) \]

Duration: \( T = 14 \) days

Area \( D \) - “France”

Figure: Temperature fluctuations

Figure: 1000 years of \( A(t) \)
Plasim: Planet Simulator, HWs in France

- **Intermediate complexity** model allows long simulation (8000 years)
- SST and the ice cover is repeated cyclically every year
- Resolution: 2.8 by 2.8 degrees. 10 vertical atmospheric levels

Figure: Plasim gridpoints

Figure: Plasim vs ERA5: return time plot

Evaluating the performance of predictions

The goal of inference: find committor function $P(Y|X)$

$$P (X = x \text{ and } Y = y) = P(x, y) = P(Y|X)P(X).$$ (5)

Evaluating the performance of predictions

The goal of inference: find committor function $P(Y|X)$

$$
\mathbb{P}(X = x \text{ and } Y = y) = P(x, y) = P(Y|X)P(X).
$$

Logarithmic (a.k.a, cross-entropy) score is suitable for rare events\(^{13}\)

$$
-S[\hat{p}_Y(X)] = - \sum_{k=0}^{K-1} Y_k \log[\hat{p}_k(x)], \quad K = 2 \text{ for binary}
$$

\(^{13}\) R. Benedetti, Monthly Weather Review (2010)
Evaluating the performance of predictions

The goal of inference: find committor function $P(Y|X)$

$$
\mathbb{P}(X = x \text{ and } Y = y) = P(x, y) = P(Y|X)P(X).
$$

Logarithmic (a.k.a, cross-entropy) score is suitable for rare events\,[13]\]

$$
-S[\hat{p}_Y(X)] = - \sum_{k=0}^{K-1} Y_k \log [\hat{p}_k(x)], \quad K = 2 \text{ for binary}
$$

In the limit of a large dataset, we have a law of large numbers

$$
\mathbb{E}\{S[\hat{p}_Y(X)]\} = - \int dx P(x) \left( \sum_{k=0}^{K-1} p_k \log p_k - \sum_{k=0}^{K-1} p_k \log \left( \frac{p_k}{\hat{p}_k} \right) \right),
$$

Evaluating the performance of predictions

The goal of inference: find **committor function** \( P(Y|X) \)

\[
\mathbb{P}(X = x \text{ and } Y = y) = P(x, y) = P(Y|X)P(X).
\]

Logarithmic (a.k.a, cross-entropy) score is suitable for rare events\[13\]

\[
-S[\hat{p}_Y(X)] = -\sum_{k=0}^{K-1} Y_k \log [\hat{p}_k(x)], \quad K = 2 \text{ for binary}
\]

In the limit of a large dataset, we have a law of large numbers

\[
\mathbb{E} \{ S[\hat{p}_Y(X)] \} = -\int dx P(x) \left( \sum_{k=0}^{K-1} p_k \log p_k - \sum_{k=0}^{K-1} p_k \log \left( \frac{p_k}{\hat{p}_k} \right) \right),
\]

**Normalized Skill Score (NSS):** subtract climatological prediction

\[
\text{NSS} = \frac{-\sum_i \overline{p}_i \log \overline{p}_i - \mathbb{E} \{ S[\hat{p}_Y(X)] \}}{-\sum_i \overline{p}_i \log \overline{p}_i}
\]

Predicting Heat Waves (HW) with Deep Learning (DL)

Probabilistic prediction: softmax output

- **Soft-max** (sigmoid) bounds to \((0, 1)\) range \[^{14}\][^{15}\]

\[
P(Y_n = k \mid x_n, \{w_{k'}\}_{k'=0}^{K-1}) = \frac{e^{-x_n^T w_k}}{\sum_{k'=0}^{K-1} e^{-x_n^T w_{k'}}},\tag{9}
\]

\[^{14}\] J. Platt et al., Advances in large margin classifiers (1999)
\[^{15}\] C. Guo et al., (2017)
Probabilistic prediction: softmax output

- **Soft-max (sigmoid) bounds to** $(0, 1)$ range $^{[14][15]}$

$$P(Y_n = k \mid x_n, \{w_{k'}\}_{k'=0}^{K-1}) = \frac{e^{-x_n^T w_k}}{\sum_{k'=0}^{K-1} e^{-x_n^T w_{k'}}},$$  

(9)

- $Y$ - binary (0: is not HW, 1: is HW):
- HW: above 95 percentile of $A(t)$

---

$^{[14]}$ J. Platt et al., Advances in large margin classifiers (1999)

$^{[15]}$ C. Guo et al., (2017)
Predicting Heat Waves (HW) with Deep Learning (DL)

Probabilistic prediction: softmax output

- **Soft-max (sigmoid) bounds to (0, 1) range** \[^{[14]}{[15]}\]
  \[
P(Y_n = k \mid \mathbf{x}_n, \{w_{k'}\}_{k'=0}^{K-1}) = \frac{e^{-x_n^Tw_k}}{\sum_{k'=0}^{K-1} e^{-x_n^Tw_{k'}}}, \tag{9}
\]

- **\(Y\) - binary (0: is not HW, 1: is HW):**
- **HW: above 95 percentile of \(A(t)\)**
- **\(X(\tau)\) - data at time \(\tau\) preceding HW**

Probabilistic prediction: softmax output

- Soft-max (sigmoid) bounds to (0, 1) range $^{[14][15]}$

$$P \left( Y_n = k \mid x_n, \{w_{k'}\}_{k'=0}^{K-1} \right) = \frac{e^{-x_n^T w_k}}{\sum_{k'=0}^{K-1} e^{-x_n^T w_{k'}}},$$ (9)

- $Y$ - binary (0: is not HW, 1: is HW):
- HW: above 95 percentile of $A(t)$
- $X(\tau)$ - data at time $\tau$ preceding HW
  - $X_0 = t_M$ - 2m temperature, France
  - $X_1 = z_G$ - 500mbar geopotential
  - $X_2 = s_M$ - soil moisture, France

---

Probabilistic prediction: softmax output

- **Soft-max** (sigmoid) bounds to $(0, 1)$ range \[^{14,15}\]

\[
P(Y_n = k | \mathbf{x}_n, \{w_{k'}\}_{k'=0}^{K-1}) = \frac{e^{-x_n^T w_k}}{\sum_{k'=0}^{K-1} e^{-x_n^T w_{k'}}},
\]

- $Y$ - binary (0: is not HW, 1: is HW):
- HW: above 95 percentile of $A(t)$
- $X(\tau)$ - data at time $\tau$ preceding HW
  - $X_0 = t_M$ - 2m temperature, France
  - $X_1 = z_G$ - 500mbar geopotential
  - $X_2 = s_M$ - soil moisture, France

---

CNN Architecture with masking

Predicting Heat Waves (HW) with Deep Learning (DL)

(Conv 2D₀ (3 × 3 × 32) → Max Pool 2D₁ (2 × 2) → Conv 2D₂ (3 × 3 × 64) → Max Pool 2D₃ (2 × 2) → Conv 2D₄ (3 × 3 × 64) → ReLu + Flat)

Yes

No

Softmax

George Miloshevich (ENSL)
george.miloshevich@ens-lyon.fr

ML Climate 2022 13 / 20
We present the plots of NSS vs lag time $\tau$ selecting different fields. $S_M$ has long-term, while $Z_G$ has short-term information. $Z_G, S_M$ coupled together account for most of the information.

**Figure: NSS 7200 years**
NSS vs different areas and data size

- We present the plots of NSS vs lag time $\tau$
- Having less data, some global teleconnections not represented well
- In reanalysis only the data from 1950 to present is available

Figure: $z_A$

Figure: $z_G$

Figure: NSS data reduction
Committor composite maps

- We plot composite maps conditioned to 99.9 percentile of $q = q(\tau)$
- The composite map reveals tripole teleconnection pattern
- We vary $\tau$ and observe that the teleconnection pattern slightly shifts
- Investigating saliency maps is the subject of current work
Outline

1. Intro to Machine Learning (ML)
2. ML in computational Earth sciences
3. Predicting Heat Waves (HW) with Deep Learning (DL)
4. Future work and conclusions
Work in progress: Rare event algorithm

- The optimal score function for \([16]\) is related to \(P(Y|X)\) committor

\[
G_k(z_k) = \sqrt{\frac{g_k(z_k)}{g_{k-1}(z_{k-1})}}, \quad \text{where (10)}
\]

\[
g_k(z_k) := \int E[h(Z_n) \mid Z_{k+1} = z']^2 P(Z_{k+1} = z' \mid Z_k = z_k) \, dz'
\] \quad \text{(11)}

Smoothness of the committor & transfer learning

- $q = q(\tau)$ is expected to be a smoothly increase closer to the heat wave
- This property is expected to play a role in rare event algorithm \[^{[17]}\]
- We achieve this by transfer learning applied to successive $\tau$

**Figure:** Training pipeline

**Figure:** $q_{tM,z_G,s_M}$ vs transfer learning

Future work and conclusions

Work in progress: The analogue Markov chain

\[ X_{n*} = \arg\min \{d (x, X_n) \} \]
\[ \{X_n\} \]

- Promising \[^{[18]}\] in Cherney-DeVore system
- **Problem**: curse of high dimensionality (\(z_G\))
- Possible solution: Dimensionality reduction
- **Issues**: Reconstruction of localized heat waves
- Possible solution: Add committor to the autoencoder loss

---

**Figure**: Analogue method: nearest neighbors

**Figure**: Schematics of a (variational) autoencoder

---

Summary

Conclusions:
- We have discussed how ML can be used to predict HWs.
- This consisted of CNN trained on 8000 years of Plasim.
- To get appreciable skill a lot of data necessary.
- Most of the information is in soil moisture and geopotential.
- Transfer learning helps achieve smoothness of the predictions.

In progress:
- Rare event algorithm: use learned probability for importance sampling.
- Analogue method: dimensionality reduction, an alternative to CNN.
- Transfer learning: Plasim $\rightarrow$ CESM $\rightarrow$ ERA5.

Acknowledgements to the future and past collaborators:
- Freddy Bouchet
- Patrice Abry
- Pierre Borgnat
- Francesco Ragone
- Dario Lucente
- Bastien Conzian
- Alessandro Lovo
- Clement Le Priol
Future work: CESM/ERA5 transfer learning

- The goal of the project: committor function for reanalysis
  - Pretrain the CNN on 8000 years long Plasim run
  - Transfer Learning to CESM (modern model consistent with IPCC)
  - Transfer Learning to ERA5 reanalysis set (perhaps fine-tuning?)

Convolutional Neural Networks (CNNs)

- Better image processing due to fewer neurons, translation invariance

Convolutional Neural Networks (CNNs)

- Better image processing due to fewer neurons, translation invariance

Convolutional Neural Networks (CNNs)

- Better image processing due to fewer neurons, translation invariance

Convolutional Neural Networks (CNNs)

- Better image processing due to fewer neurons, translation invariance
- CNNs achieve state-of-the-art results on many benchmark datasets[21]

A. Krizhevsky et al., Advances in neural information processing systems (2012)

George Miloshevich (ENSL) george.miloshevich@ens-lyon.fr