Probabilistic forecasting of heat waves with deep learning

G. Miloshevich [1]

ENS DE LYON

[1]Departement de Physique
Ecole Normale Superieure de Lyon

IXXI

Institut Rhônalpin des systèmes complexes

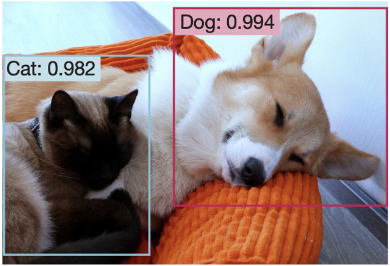
cnrs

Machine Learning and sampling methods
for climate and physics, 2022

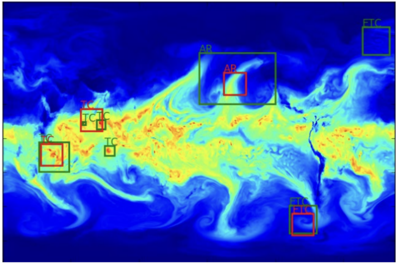
Machine Learning (ML) for extreme events

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

Object classification and localization



Pattern classification



[1] S. Seneviratne et al., Climate Change 2021: Sixth Assessment Report of the IPCC ()
[2] E. Racah et al., Advances in Neural Information Processing Systems (2017)
[3] V. Jacques-Dumas et al., Frontiers in Climate (2022)

Outline

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)

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

Outline

ANNs: image, speach recognition, games

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

[4] P. Mehta et al., Physics Reports (2019)
[5] D. E. Rumelhart et al., Nature (1986)
[6] G. Cybenko, Mathematics of Control, Signals and Systems (1989)

ANNs: image, speach recognition, games

- ML consists of various fields: [4]
  - Supervised learning
  - Unsupervised learning
  - Reinforecement learning
- Components of ANNs:
  - Hyperparameters $\boldsymbol{\theta}$, e.g. weights $\boldsymbol{w}_i$
  - Nonlinear activation function
  - loss function $E(\boldsymbol{\theta}) = C(\boldsymbol{X}, \boldsymbol{g}(\boldsymbol{\theta}))$
  - backprogapation to minimize loss [5]

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla_\theta \sum_{i \in B_k} e_i \left(\boldsymbol{X}_i, \boldsymbol{\theta}\right) \quad (1)$$

  - Universal function approximators[6]

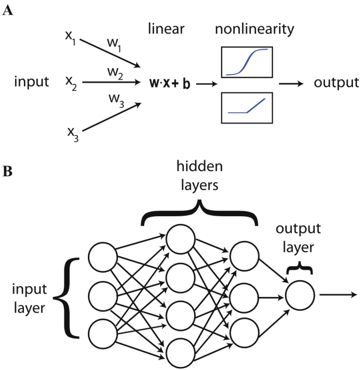


Figure: architecture

[4] P. Mehta et al., Physics Reports (2019)
[5] D. E. Rumelhart et al., Nature (1986)
[6] G. Cybenko, Mathematics of Control, Signals and Systems (1989)

Outline

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

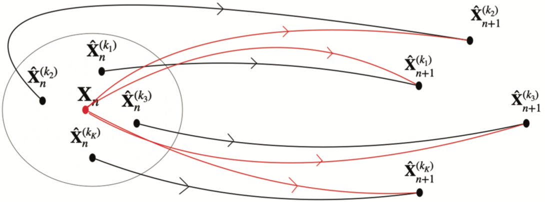


Figure: Analogue method

[7] E. N. Lorenz, Journal of Atmospheric Sciences (1969)

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]

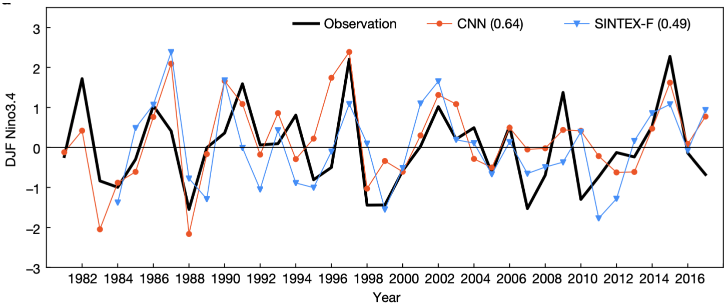


Figure: Nino3.4 indexes for an 18-month-lead

[8] Y.-G. Ham et al., Nature (2019)

[9] V. Balaji, Phil. Trans.of the Royal Soc.A: Math., Phys.and Eng. Sciences (2021)

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

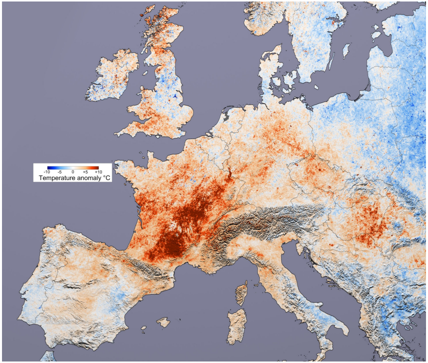


Figure: European heat wave 2003

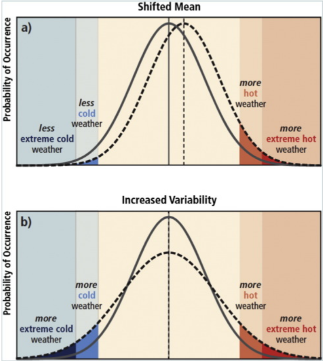


Figure: Changes in temperatures[11]

[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

Scandinavian blocking: HW onset

- Rossby wave breaking and blocking
- Advection: persistent anticyclonic anomaly

$$z(p) = R \int_p^{p_s} \frac{T}{g} \frac{dp}{p} \quad (3)$$

$$\underbrace{V = \frac{\hat{k}}{f}}_{\text{Coriolis parameter}} \times \underbrace{\nabla z}_{\text{500 mbar geopotential height}} \quad (2)$$

- Dry soil contributes to heating due to lack of latent heat

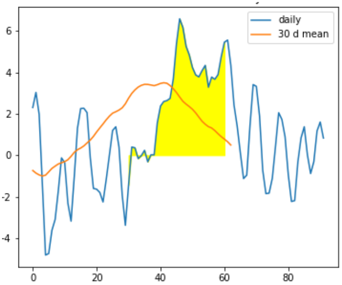


Figure: Scandinavia: Average temperature

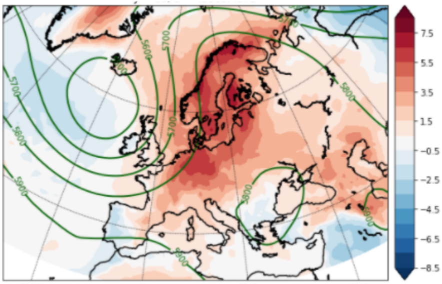


Figure: Temperature, geopotential (ECMWF)

Summer HWs over France: definition

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

$$A_T(t) = \frac{1}{T}\int_t^{t+T} \frac{1}{|\mathcal{D}|}\int_D \left(T_{2m} - \mathbb{E}\left(T_{2m}\right)\right)(\vec{r},u)\,\mathrm{d}\vec{r}\mathrm{d}u \tag{4}$$

Duration: $T = 14$ days

Area D - "France"

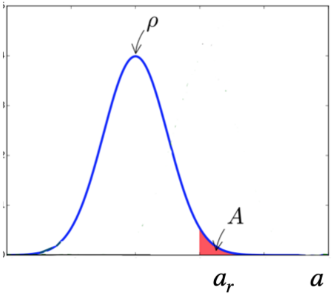


Figure: Temperature fluctuations

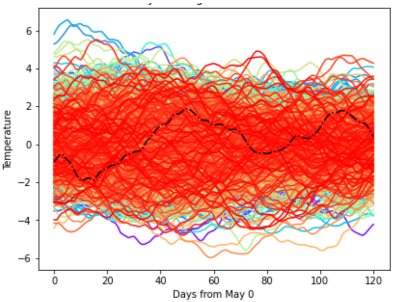


Figure: 1000 years of $A(t)$

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

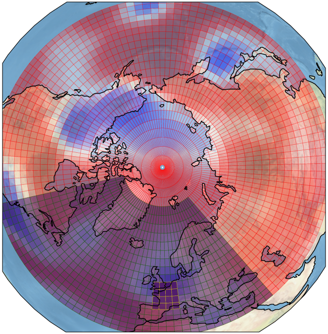


Figure: Plasim gridpoints

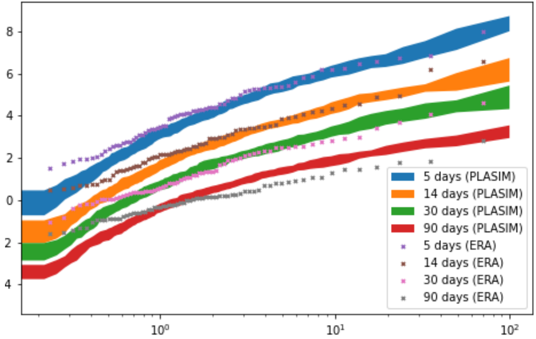


Figure: Plasim vs ERA5: return time plot [12]

[12] G. Miloshevich et al., (Apr. 2021)

Evaluating the performance of predictions

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

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

[13] R. Benedetti, Monthly Weather Review (2010)

Evaluating the performance of predictions

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

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

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

$$-S\left[\hat{p}_Y(X)\right] = -\sum_{k=0}^{K-1} Y_k \log\left[\hat{p}_k(x)\right], \quad K = 2 \text{ for binary} \tag{6}$$

[13] R. Benedetti, Monthly Weather Review (2010)

Evaluating the performance of predictions

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

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

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

$$-S\left[\hat{p}_Y(X)\right] = -\sum_{k=0}^{K-1} Y_k \log\left[\hat{p}_k(x)\right], \quad K = 2 \text{ for binary} \tag{6}$$

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

$$\mathbb{E}\left\{S\left[\hat{p}_Y(X)\right]\right\} = -\int \mathrm{d}x P(x)\left(\sum_{k=0}^{K-1} p_k \log p_k - \sum_{k=0}^{K-1} p_k \log\left(\frac{p_k}{\hat{p}_k}\right)\right), \tag{7}$$

[13] R. Benedetti, Monthly Weather Review (2010)

Evaluating the performance of predictions

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

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

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

$$-S\left[\hat{p}_Y(X)\right] = -\sum_{k=0}^{K-1} Y_k \log\left[\hat{p}_k(x)\right], \quad K = 2 \text{ for binary} \tag{6}$$

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

$$\mathbb{E}\left\{S\left[\hat{p}_Y(X)\right]\right\} = -\int \mathrm{d}x P(x) \left(\sum_{k=0}^{K-1} p_k \log p_k - \sum_{k=0}^{K-1} p_k \log\left(\frac{p_k}{\hat{p}_k}\right)\right), \tag{7}$$

Normalized Skill Score (NSS): subtract climatological prediction

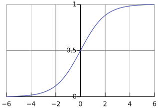
$$\text{NSS} = \frac{-\sum_i \overline{p}_i \log \overline{p}_i - \mathbb{E}\left\{S\left[\hat{p}_Y(X)\right]\right\}}{-\sum_i \overline{p}_i \log \overline{p}_i} \tag{8}$$

[13] R. Benedetti, Monthly Weather Review (2010)

Probabilistic prediction: softmax output

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

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

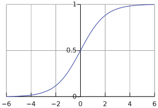


[14] J. Platt et al., Advances in large margin classifiers (1999)
[15] C. Guo et al., (2017)

Probabilistic prediction: softmax output

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

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



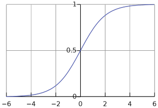
- Y - binary (0: is not HW, 1: is HW):
- HW: above 95 percentile of $A(t)$

[14] J. Platt et al., Advances in large margin classifiers (1999)
[15] C. Guo et al., (2017)

Probabilistic prediction: softmax output

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

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



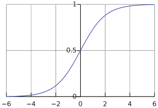
- Y - binary (0: is not HW, 1: is HW):
- HW: above 95 percentile of $A(t)$
- $X(\tau)$ - data at time τ preceding HW

[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]

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



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

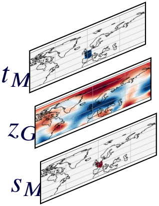


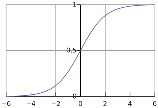
Figure: Possible field inputs

[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]

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



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

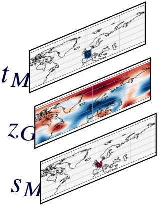
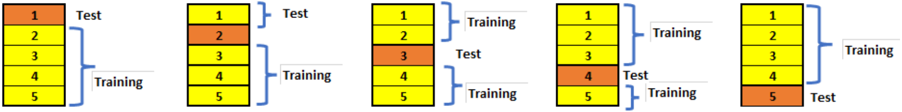
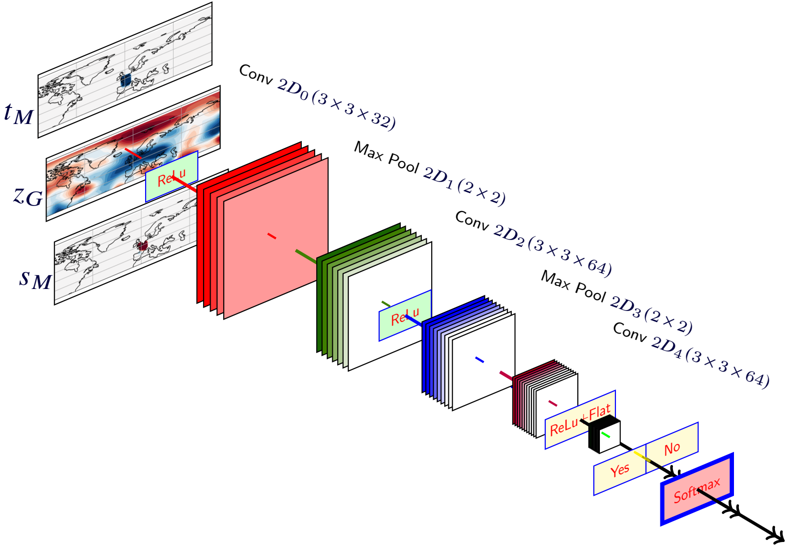


Figure: Possible field inputs



[14] J. Platt et al., Advances in large margin classifiers (1999)
[15] C. Guo et al., (2017)

CNN Architecture with masking



NSS vs lag time for different fields

- We present the plots of NSS vs lag time τ 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

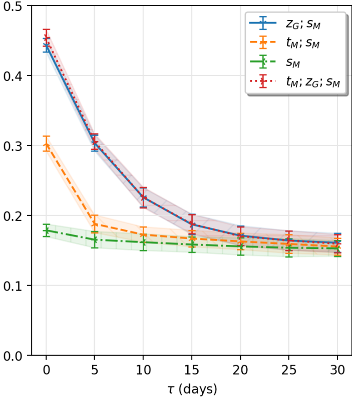


Figure: NSS 7200 years

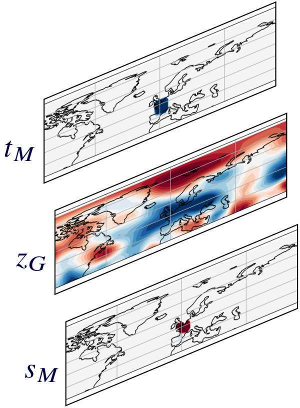


Figure: Possible field inputs

NSS vs different areas and data size

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

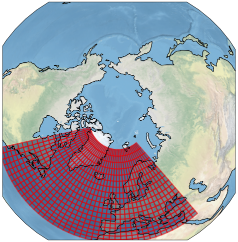


Figure: z_A

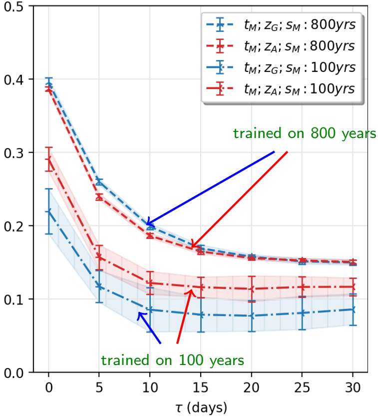


Figure: NSS data reduction

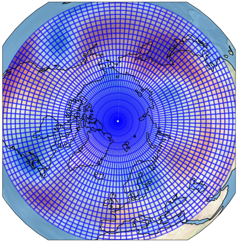


Figure: z_G

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 τ and observe that the teleconnection pattern slightly shifts
- Investigating saliency maps is the subject of current work

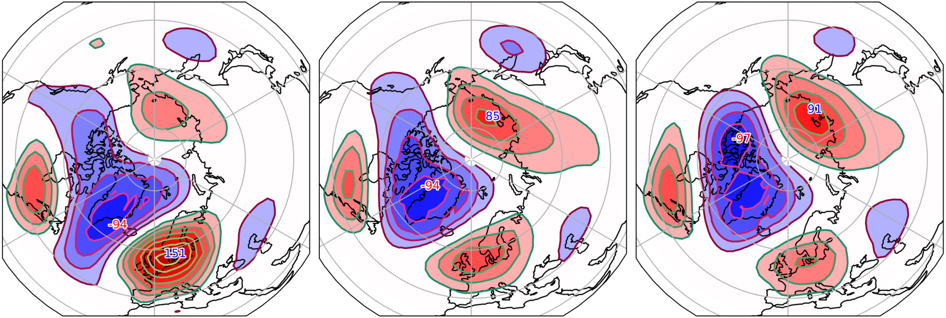


Figure: $\tau = 0$ Figure: $\tau = -5$ Figure: $\tau = -10$

Outline

Work in prograss: Rare event algorithm

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

$$G_k\left(\boldsymbol{z}_k\right)=\sqrt{\frac{g_k\left(\boldsymbol{z}_{\boldsymbol{k}}\right)}{g_{k-1}\left(\boldsymbol{z}_{\boldsymbol{k-1}}\right)}}, \qquad \text{where} \quad (10)$$

$$g_k(\boldsymbol{z}_k) := \int \mathbb{E}\left[h\left(\boldsymbol{Z}_n\right) \mid Z_{k+1}=z'\right]^2 P\left(Z_{k+1}=z' \mid Z_k=z_k\right) dz' \quad (11)$$

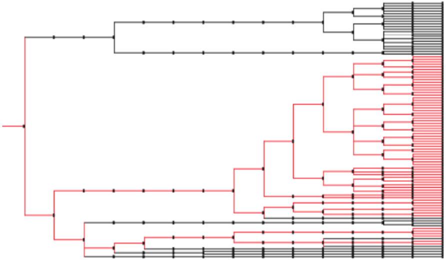


Figure: Geneological rare event algorithm

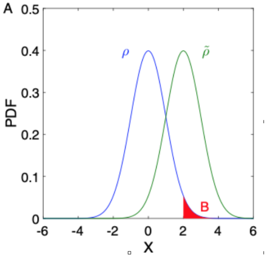


Figure: Importance sapling

[16] H. Chraibi et al., Monte Carlo Methods and Applications (2021)

Smoothness of the committor & transfer learning

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

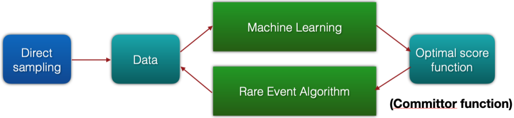


Figure: Training pipeline

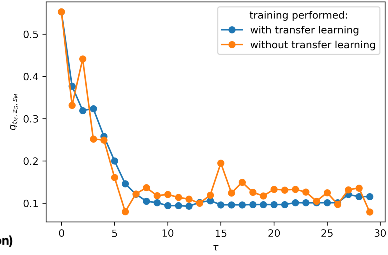


Figure: q_{t_M, z_G, s_M} vs transfer learning

[17] F. Ragone and F. Bouchet, Geophysical Research Letters (2021)

Work in prograss: The analogue Markov chain

$$X_{n_\star} = \underset{\{X_n\}}{\operatorname{argmin}} \{d(x, X_n)\}$$

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

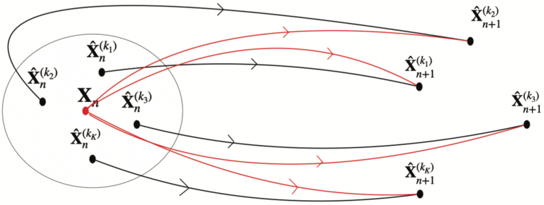


Figure: Analogue method: nearest neighbors

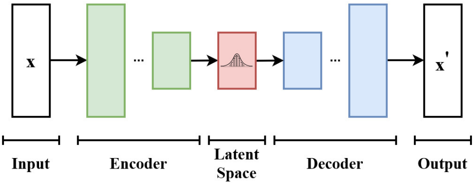


Figure: Schematics of a (variational) autoencoder

[18] D. Lucente et al., arXiv preprint arXiv:2110.05050 (2021)

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 $\longrightarrow$ CESM $\longrightarrow$ ERA5

Ackwnoledgements to the future and past collaborators:

- Freddy Bouchet
- Patrice Abry
- Pierre Borgnat
- Francesco Ragone
- Dario Lucente
- Bastien Conzian
- Alessandro Lovo
- Clement Le Priol

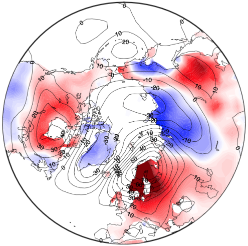


Figure: Plasim rare event[19]

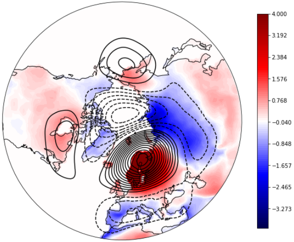


Figure: CESM composite[20]

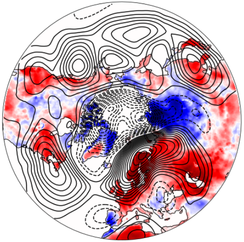


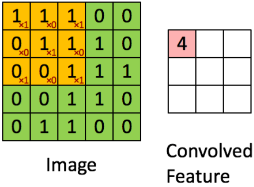
Figure: ERA5 July 2018

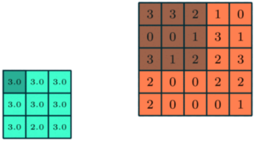
- The goal of the project: committor function for reanalysis
  - Pretrain the CNN on 8000 years long Plasim run
  - Transfer Learning to CESM (modern model conistent with IPCC)
  - Transfer Learning to ERA5 reanalysis set (perhaps fine-tuning?)

[19] F. Ragone et al., Proceedings of the National Academy of Sciences (2018)
[20] G. Miloshevich et al., "Drivers of midlatitude extreme heat waves revealed by analogues and machine learning", in Egu general assembly conference abstracts, EGU General Assembly Conference Abstracts (Apr. 2021), EGU21–15642

Convolutional Neural Networks (CNNs)

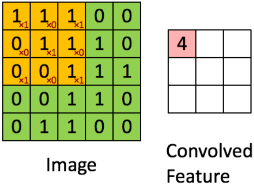
- Better image processing due to fewer neurons, translation invariance

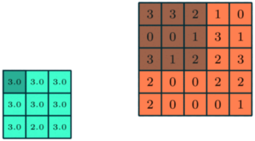




[21] A. Krizhevsky et al., Advances in neural information processing systems (2012)

- Better image processing due to fewer neurons, translation invariance

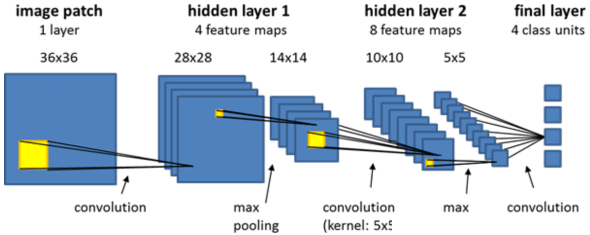


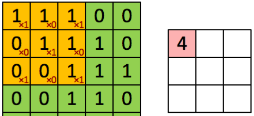


[21] A. Krizhevsky et al., Advances in neural information processing systems (2012)

Convolutional Neural Networks (CNNs)

- Better image processing due to fewer neurons, translation invariance





Image

Convolved Feature

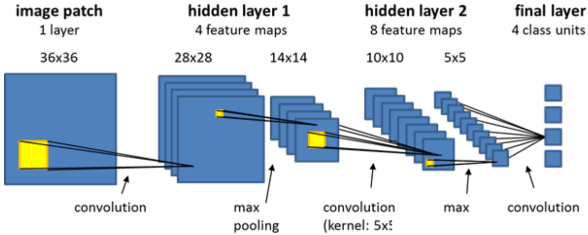
3.0	3.0	3.0
3.0	3.0	3.0
3.0	2.0	3.0

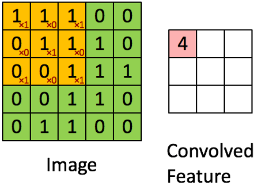
3	3	2	1	0
0	0	1	3	1
3	1	2	2	3
2	0	0	2	2
2	0	0	0	1

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

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]





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