

An aerial photograph of a town, likely in the Alps, is shown from a high angle. The town is surrounded by green hills and is partially obscured by thick, white clouds. Overlaid on the bottom half of the image is a weather map with white contour lines and arrows indicating wind direction and speed. The map shows pressure contours ranging from 1010 to 1025 hPa. The background of the entire slide is a deep blue sky with wispy white clouds.

Utilisation du deep learning afin de renormaliser la matrice de covariance basée sur l'opérateur de diffusion dans NEMOVar.

Vincent Chabot (GMAP/ASSIM) and Marcin Chrust (ECMWF)

February 6, 2024



METEO FRANCE

Assimilation de données

Trouver x minimisant

$$J(x) = \|x - x^b\|_{\mathbf{B}^{-1}} + \|y^o - \mathcal{H}(x)\|_{\mathbf{R}^{-1}}$$

avec $\|x\|_{\mathbf{B}^{-1}} = x^T \mathbf{B}^{-1} x$.

Dans notre cas, la matrice \mathbf{B} est décomposée en :

$$\mathbf{B} = \mathbf{D}^{1/2} \mathbf{C} \mathbf{D}^{1/2}$$

$$(\mathbf{C}^{1/2})^T = (\mathbf{C}^T)^{1/2}$$

$$\mathbf{C}^{1/2} = \begin{pmatrix} \mathbf{C}_S^{1/2} & 0 & 0 \\ 0 & \mathbf{C}_T^{1/2} & 0 \\ 0 & 0 & \mathbf{C}_\eta^{1/2} \end{pmatrix}$$

avec \mathbf{C} une matrice de corrélation.

Définition du problème

[Weaver and Courtier, 2001] La matrice de corrélation peut être obtenue en résolvant l'équation différentielle suivante :

$S(z)$: Signal d'entrée (incrément)

$$\eta(0, z) = S(z)$$
$$\frac{\partial \eta(t, z)}{\partial t} - \nabla \cdot (\kappa(z) \nabla \eta(t, z)) = 0$$

Le signal filtré $\eta(t, z) = \mathbf{F}^{1/2} S(z)$ est alors proportionnel à $\mathbf{C}^{1/2} S(z)$.

But du travail

Déterminer le coefficient de renormalisation, c'est à dire, déterminer la matrice diagonale Λ telle que

$$\mathbf{C} = \Lambda \mathbf{F} \Lambda$$

Comment cela est-il fait aujourd'hui ?

- ▶ Par une méthode exacte en utilisant $F\delta_{xx}$ pour tous les points de la grille
- ▶ En utilisant des tirages aléatoires (10 000 membres).

De manière opérationnel, le CEP utilise les tirages aléatoires (12h sur HPC).

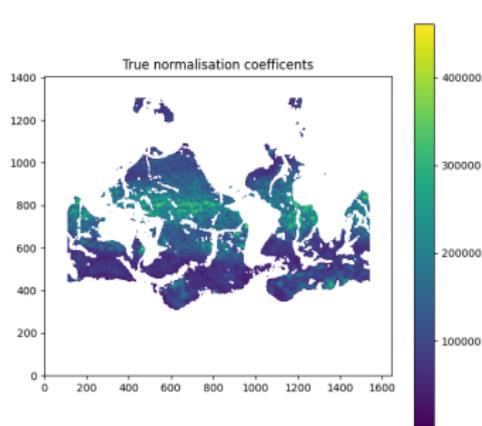
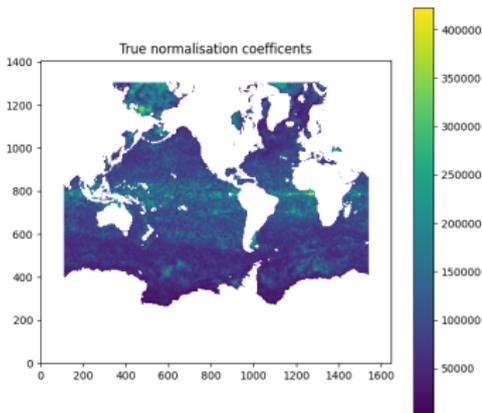
Trop cher dans un contexte opérationnel \Rightarrow Une matrice climatologique est utilisé. Pour l'atmosphère, on a vu qu'il y avait des bénéfices à utiliser des matrices de covariances "du jour".

Utiliser du Deep Learning pour cette tâche particulière n'est pas complètement nouveau [Skrunes, 2022] (CERFACS) et les travaux de **J. While** (MetOffice).

Données disponibles

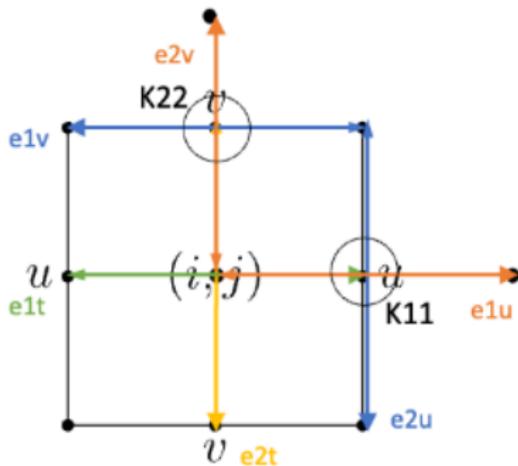
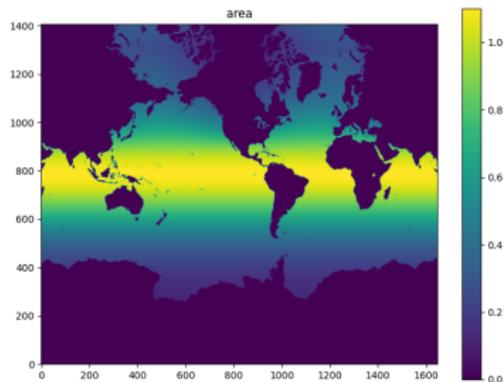
- ▶ Coefficient de diffusion (estimés sur l'ensemble (10 membres) du jour)
- ▶ Coefficient de renormalisation : Tirage aléatoire de 10 000 perturbations.
- ▶ Information sur la grille (profondeur, aire, distance à la côte)

Grille : Orca 0.25 grid (1207 × 1242) avec 75 niveaux. Grille irrégulière Tri-polaire. Masque Terre/Mer évoluant suivant les niveaux.



5 ans de données, à raison d'un échantillon tous les 5 jours.

Pré-traitement : Padding strategy et renormalisation "Physique"



From [Skrunes, 2022]

- ▶ Coefficient de diffusion : $\kappa_{11} = \kappa_{11} * e2u/e1u$
- ▶ Coefficient de diffusion : $\kappa_{22} = \kappa_{22} * e1v/e2v$
- ▶ Racine carré du coefficient de normalisation : $\gamma_u = \gamma/\sqrt{e3t}$

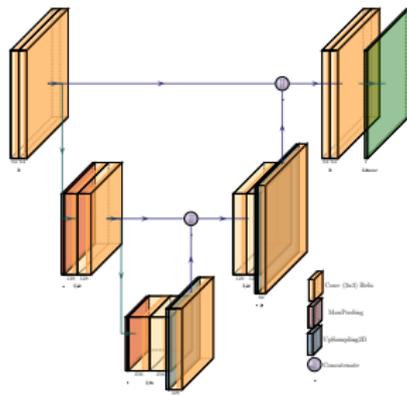
Entraînement

Entrées : κ_{11} , κ_{22} , Aire, Distance à la côte, épaisseur de la couche

Sortie : γ

Réseau : Type Unet

Métrique utilisée : $\frac{|\gamma^2 - \tilde{\gamma}^2|}{\gamma^2}$

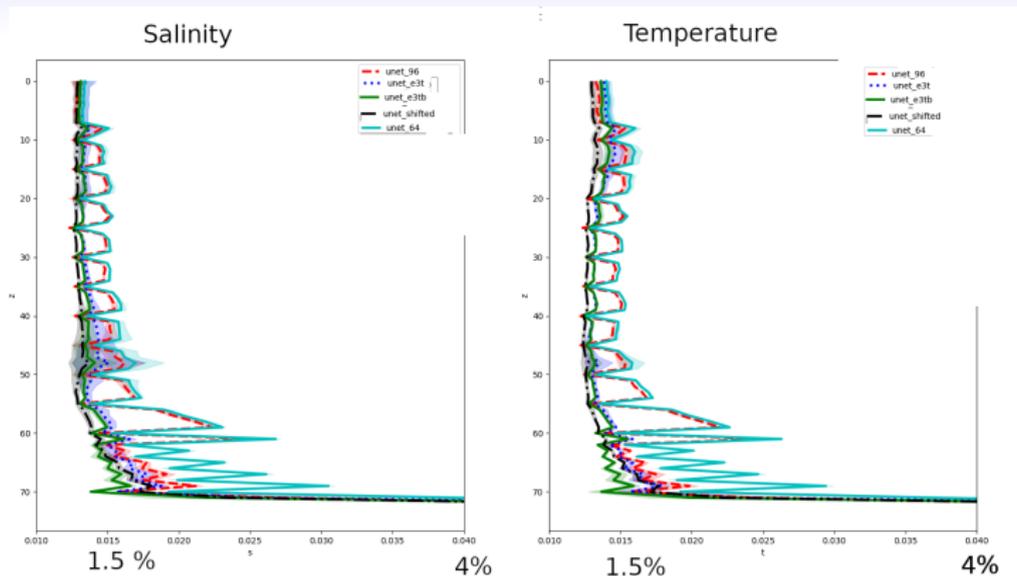


Niveau : Chaque niveau constitue un exemple différent (problème 2D).
On utilise dans le jeu d'entraînement 1 niveau tous les 5 jusqu'au niveau 60 puis chaque niveau après le 62.

Période d'entraînement : 2012- 2015

Test : Début 2016 (6 mois) **Validation :** Fin 2016 (6 mois)

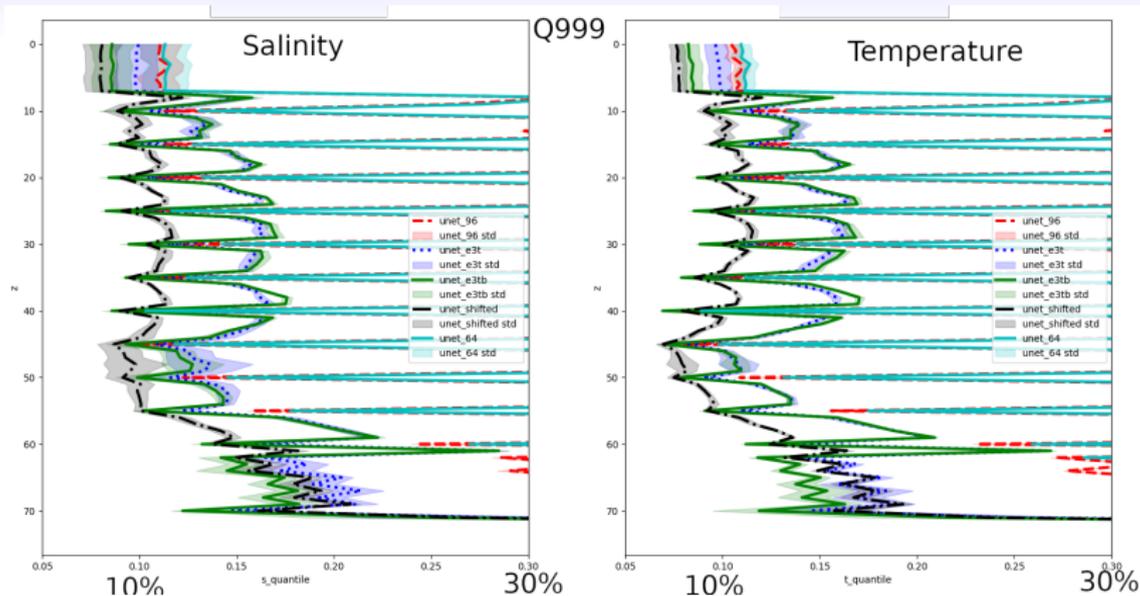
Erreur moyenne par niveau (Validation)



cyan et **rouge** n'incluent pas l'épaisseur de la couche. **cyan** : un niveau sur 2 après 60.

Noir : Données d'entraînements shiftées de manière aléatoire ($[-5; 5]$)

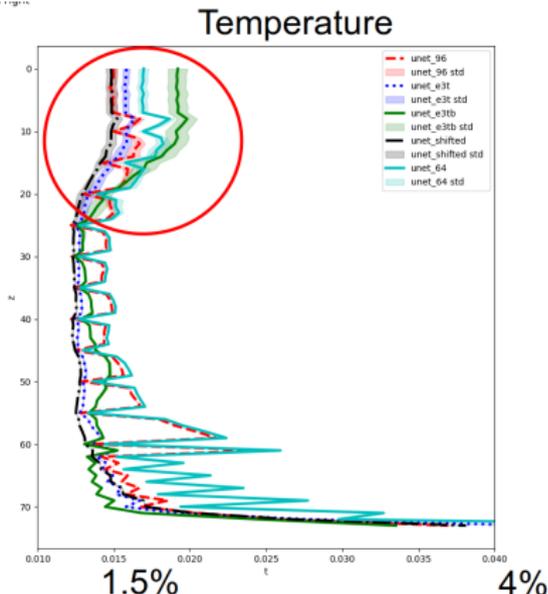
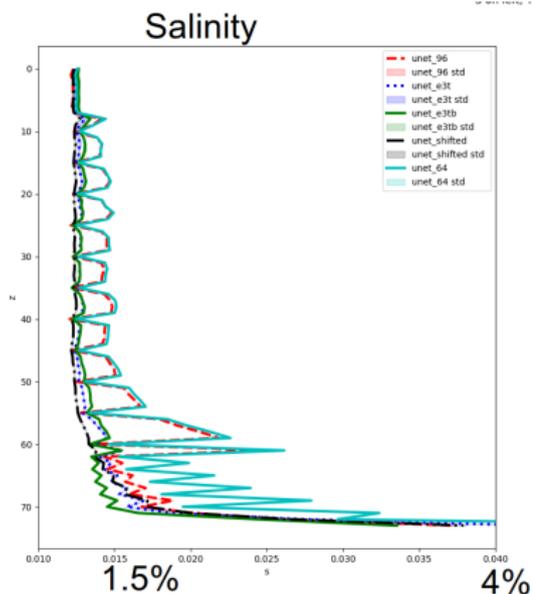
Q 999



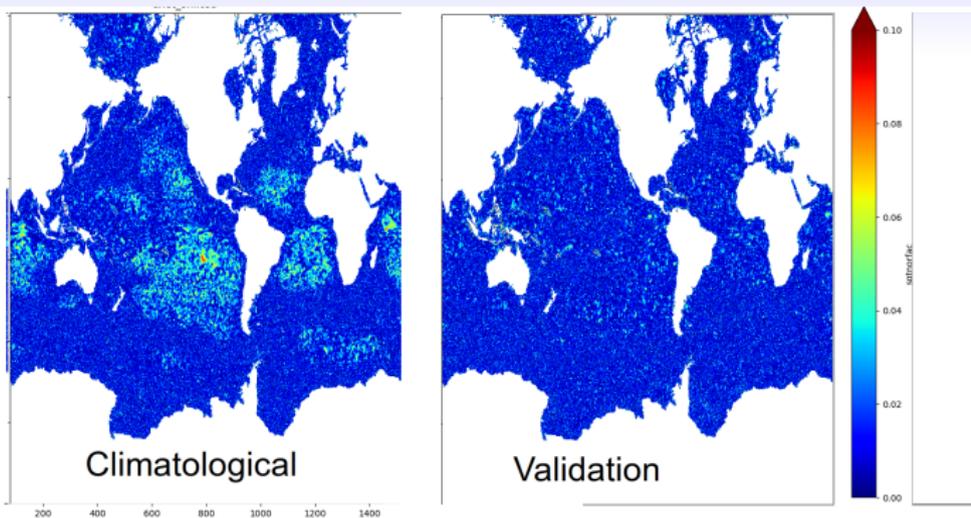
cyan et **rouge** n'incluent pas l'épaisseur de la couche. **cyan** : un niveau sur 2 après 60.

Noir : Données d'entraînements shiftées de manière aléatoire ($[-5; 5]$)

Vérification sur le dataset climatologique



Vérification sur le dataset climatologique



Beaucoup d'erreur au niveau de l'équateur.

Bonne nouvelle : Les quantiles les plus importants ne sont pas affectés.

Tentative d'explication : Changement de distribution des coefficients de diffusion entre les deux datasets : coefficients plus faible (et plus lisse) pour le dataset climatologique.

Et si on utilisait des méthodes simples ?

Les erreurs dépendent de la position sur la grille : pour chaque niveau, un petit nombre de pixels ont une erreur importante.

Peut-on dans ce cadre utiliser une correction de biais ?

- ▶ Correction de biais additive
- ▶ Correction de biais multiplicative

Biais est calculé sur le dataset de test.

Et si on utilisait des méthodes simples ?

Les erreurs dépendent de la position sur la grille : pour chaque niveau, un petit nombre de pixels ont une erreur importante.

Peut-on dans ce cadre utiliser une correction de biais ?

- ▶ Correction de biais additive
- ▶ Correction de biais multiplicative

Biais est calculé sur le dataset de test.

Erreur moyenne

	SSH	Salin.	Temp
Original	1.80	1.32	1.31
Additive	0.77	0.61	0.66
Mult.	0.65	0.56	0.54

Dataset de validation 😊

Et si on utilisait des méthodes simples ?

Les erreurs dépendent de la position sur la grille : pour chaque niveau, un petit nombre de pixels ont une erreur importante.

Peut-on dans ce cadre utiliser une correction de biais ?

- ▶ Correction de biais additive
- ▶ Correction de biais multiplicative

Biais est calculé sur le dataset de test.

Erreur moyenne

	SSH	Salin.	Temp
Original	1.80	1.32	1.31
Additive	0.77	0.61	0.66
Mult.	0.65	0.56	0.54

Dataset de validation 😊

Erreur moyenne

	SSH	Salin.	Temp.
Original	1.49	1.26	1.34
Additive	1.0	1.70	1.76
Mult.	0.87	1.62	1.68

Dataset climatologique 😞

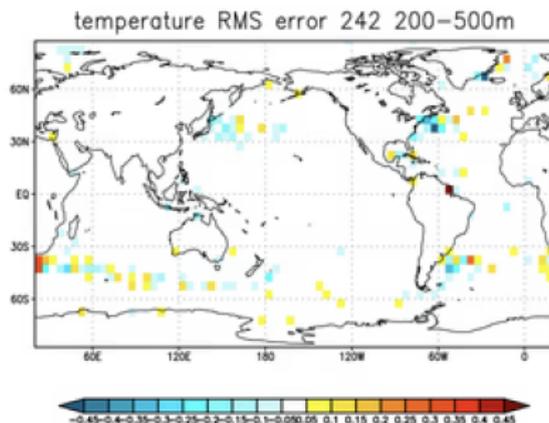
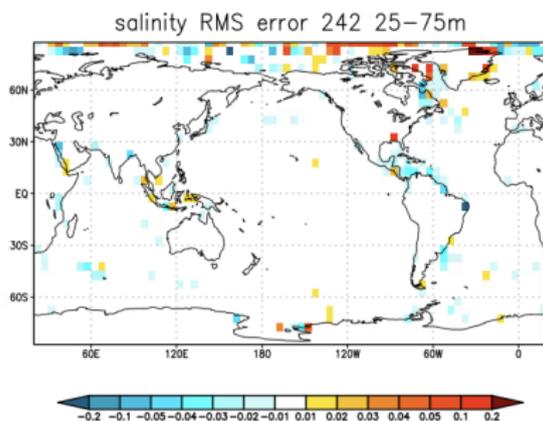
Impact sur l'assimilation

Combien de temps cela prend-il pour tester une solution ?

Deep learning : quelques heures sur un GPU.

En assimilation : quelques mois sur HPC.

L'impact des différents réseau est similaire (avec ou sans correction de biais).



Conclusion

- ▶ Bonne estimation moyenne des coefficients de normalisation en utilisant du Deep Learning
- ▶ Des erreurs importantes (liées à la géométrie et à la distribution des coefficients de diffusion) sont toujours présentes,
- ▶ Des expériences d'impact ont été conduites. Il reste du travail à faire sur la matrice **B**.

[Skrunes, 2022] **Skrunes, F. (2022).**

Deep learning methods for estimating normalisation coefficients of diffusion operators.

Master's thesis, INSA.

[Weaver and Courtier, 2001] **Weaver, A. and Courtier, P. (2001).**

Correlation modelling on the sphere using a generalized diffusion equation.

Quarterly Journal of the Royal Meteorological Society, 127(575):1815–1846.