

Prédiction rapide d'inondations de tsunamis par réseaux de neurones

Tom Sprunck

CEA - DASE - SLDG - LEGA

26 août 2021

1. Introduction
 1. Contexte
 2. Cadre
 3. Facteurs étudiés
2. Bases de données
 1. Génération des simulations
 2. Découpage des données
3. Choix des architectures
 1. Paramétrage d'un MLP
 2. Création d'un CNN
 3. Comparaison des modèles
4. Évaluation
 1. Sélection et évaluation d'un modèle
 2. Généralité du modèle
5. Conclusion

Sommaire

1. Introduction

1. Contexte

2. Cadre

3. Facteurs étudiés

2. Bases de données

3. Choix des architectures

4. Évaluation

5. Conclusion

Commissariat à l'Énergie Atomique et aux Énergies Alternatives :

- ▶ Etablissement Public à caractère Industriel et Commercial (EPIC)
- ▶ 9 centres d'études civiles et militaires
- ▶ 4 domaines : défense et sécurité, énergie, recherche pour l'industrie et recherche fondamentale en sciences de la matière et sciences de la vie

Commissariat à l'Énergie Atomique et aux Énergies Alternatives :

- ▶ Etablissement Public à caractère Industriel et Commercial (EPIC)
- ▶ 9 centres d'études civiles et militaires
- ▶ 4 domaines : défense et sécurité, énergie, recherche pour l'industrie et recherche fondamentale en sciences de la matière et sciences de la vie

↳ **Centre National d'Alerte aux Tsunamis (CENALT) :**

- ▶ surveillance du risque tsunami sur les côtes françaises en Méditerranée et en Atlantique nord-est
- ▶ prévision d'occurrences de tsunamis d'origine sismique
- ▶ estimation des temps d'arrivée et des amplitudes à la côte

Enoncé du problème

Problème : alerter sur les potentiels tsunامي en méditerranée en moins de 15 minutes après un séisme

Un des enjeux : calculer les hauteurs de vagues à la côte et les inondations après la propagation du tsunami

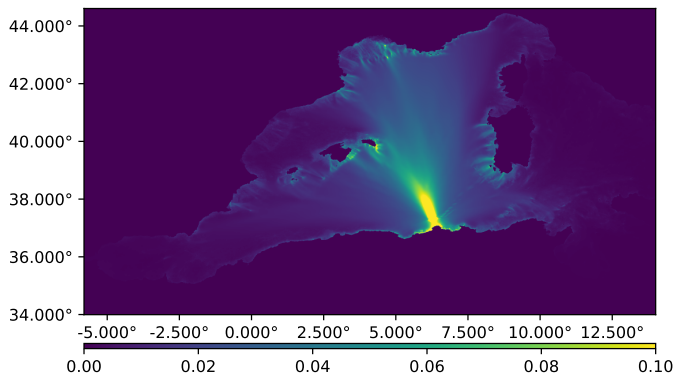


Figure: Simulation pour un séisme de magnitude 6.6

Méthode

Objectif : prédire les hauteurs de vagues et les inondations à la côte en haute résolution

Méthode

Objectif : prédire les hauteurs de vagues et les inondations à la côte en haute résolution

→ simulation numérique multi-grilles (grilles imbriquées à résolution croissante)

Méthode

Objectif : prédire les hauteurs de vagues et les inondations à la côte en haute résolution

→ simulation numérique multi-grilles (grilles imbriquées à résolution croissante)

Problème : temps de calcul trop élevé en contexte d'alerte

→ développement de méthodes de prédiction rapide

Méthode

Objectif : prédire les hauteurs de vagues et les inondations à la côte en haute résolution

→ simulation numérique multi-grilles (grilles imbriquées à résolution croissante)

Problème : temps de calcul trop élevé en contexte d'alerte

→ développement de méthodes de prédiction rapide

Traitement IA :

- ▶ calcul des hauteurs maximales sur la grille la plus grossière (< 1 minute)
- ▶ prédiction des hauteurs maximales sur la grille fine avec un réseau de neurones (< 1s)

→ prévision rapide des hauteurs max sur la grille fine

Facteurs étudiés

- ▶ choix des grilles (taille, localisation, résolution...)

Facteurs étudiés

- ▶ choix des grilles (taille, localisation, résolution...)
- ▶ prétraitement des entrées/sorties (normalisation...)

Facteurs étudiés

- ▶ choix des grilles (taille, localisation, résolution...)
- ▶ prétraitement des entrées/sorties (normalisation...)
- ▶ choix de l'architecture du réseau

Facteurs étudiés

- ▶ choix des grilles (taille, localisation, résolution...)
- ▶ prétraitement des entrées/sorties (normalisation...)
- ▶ choix de l'architecture du réseau
- ▶ choix des hyperparamètres du modèle

Facteurs étudiés

- ▶ choix des grilles (taille, localisation, résolution...)
- ▶ prétraitement des entrées/sorties (normalisation...)
- ▶ choix de l'architecture du réseau
- ▶ choix des hyperparamètres du modèle
- ▶ régularisation, surapprentissage du modèle

Facteurs étudiés

- ▶ choix des grilles (taille, localisation, résolution...)
- ▶ prétraitement des entrées/sorties (normalisation...)
- ▶ choix de l'architecture du réseau
- ▶ choix des hyperparamètres du modèle
- ▶ régularisation, surapprentissage du modèle
- ▶ **les données** (quantité, variété, réalisme ...)

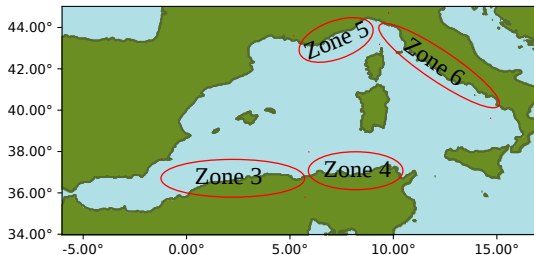
Sommaire

1. Introduction
2. Bases de données
 1. Génération des simulations
 2. Découpage des données
3. Choix des architectures
4. Évaluation
5. Conclusion

Génération des simulations

Les séismes considérés proviennent de quatre zones :

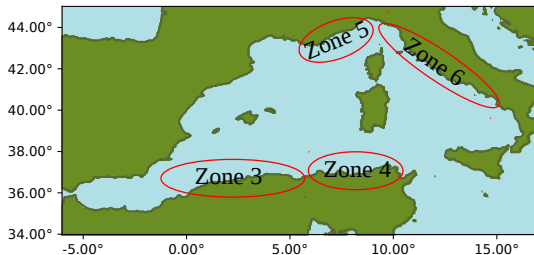
- ▶ nord de l'Algérie (z03), nord de la Tunisie (z04)
- ▶ mer Ligure (z05)
- ▶ ouest de l'Italie (z06)



Génération des simulations

Les séismes considérés proviennent de quatre zones :

- ▶ nord de l'Algérie (z03), nord de la Tunisie (z04)
- ▶ mer Ligure (z05)
- ▶ ouest de l'Italie (z06)



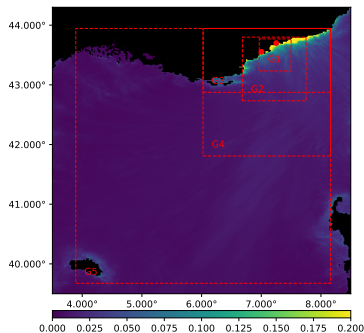
- ▶ les paramètres de source proviennent de la base de faille du CENALT et d'une étude probabiliste du risque tsunami
- ▶ on calcule les simulations à l'aide du code Taitoko développé au CEA

- ▶ 3 bases de données générées : Nice (≈ 1600 scénarios), Cannes et Antibes (≈ 400 scénarios)
- ▶ 10% des scénarios servent de jeu de test
- ▶ 20% des scénarios restants servent de jeu de validation
- ▶ on extrait une même proportion de chaque zone pour garder des jeux homogènes

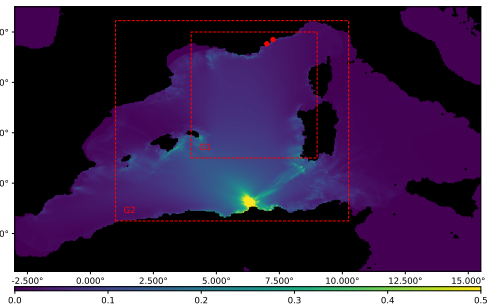
- ▶ 3 bases de données générées : Nice (≈ 1600 scénarios), Cannes et Antibes (≈ 400 scénarios)
- ▶ 10% des scénarios servent de jeu de test
- ▶ 20% des scénarios restants servent de jeu de validation
- ▶ on extrait une même proportion de chaque zone pour garder des jeux homogènes
- ▶ on garde les grilles basse (mère) et haute (fille) résolution de chaque simulation
- ▶ on extrait des sous-grilles des grilles mères et filles pour les prédictions

Grilles mères

Résolutions : 3700m pour Cannes, 900m pour Nice/Antibes

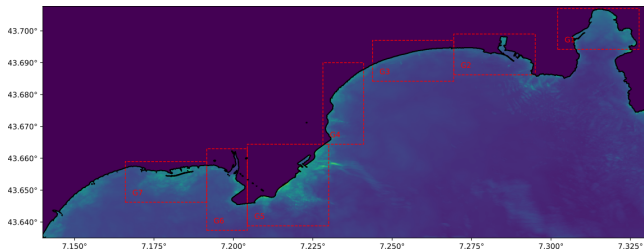


(a) Grilles mères pour Nice

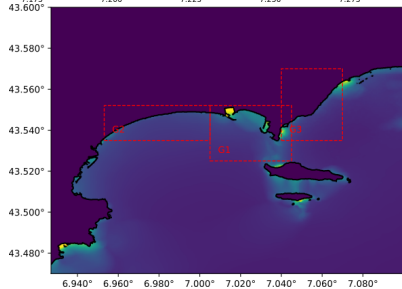


(b) Grilles mères pour Cannes

Grilles filles, résolution à 10m



(a) Nice



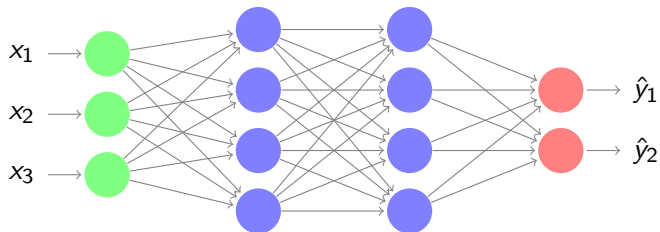
(b) Cannes

Sommaire

1. Introduction
2. Bases de données
3. Choix des architectures
 1. Paramétrage d'un MLP
 2. Création d'un CNN
 3. Comparaison des modèles
4. Évaluation
5. Conclusion

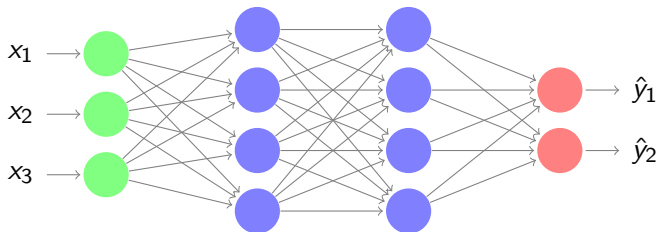
Réseau MLP

Premier réseau : MLP classique



Réseau MLP

Premier réseau : MLP classique



Plusieurs paramètres à ajuster :

- ▶ le nombre de couches
- ▶ le nombre de neurones par couche
- ▶ la régularisation (taux de dropout, pénalisation L^2)
- ▶ les paramètres d'entraînement (taux d'apprentissage...)

Réduction de dimension

Les vecteurs d'entrée et de sortie sont de taille importante

↔ on peut chercher à réduire leur dimension

Réduction de dimension

Les vecteurs d'entrée et de sortie sont de taille importante

↔ on peut chercher à réduire leur dimension

On peut supprimer des cellules selon la profondeur :

- ▶ en entrée ignorer les cellules de profondeur nulle
- ▶ en sortie considérer uniquement les cellules d'une tranche de profondeur fixée

Réduction de dimension

Les vecteurs d'entrée et de sortie sont de taille importante

↪ on peut chercher à réduire leur dimension

On peut supprimer des cellules selon la profondeur :

- ▶ en entrée ignorer les cellules de profondeur nulle
- ▶ en sortie considérer uniquement les cellules d'une tranche de profondeur fixée

Deux méthodes de réduction testées pour la dimension d'entrée :

- ▶ analyse en composantes principales (ACP)
- ▶ réseau auto-encodeur simple

Recherche d'hyperparamètres sur Cannes

Les trois paramètres ajustés sont :

- ▶ le nombre de couches cachées
- ▶ le nombre de neurones par couche
- ▶ le taux d'apprentissage

Recherche d'hyperparamètres sur Cannes

Les trois paramètres ajustés sont :

- ▶ le nombre de couches cachées
- ▶ le nombre de neurones par couche
- ▶ le taux d'apprentissage

La méthode de sélection est la suivante :

- ▶ on teste plusieurs niveaux de réduction en entrée (aucune, 50, 100, 125 composantes)

Recherche d'hyperparamètres sur Cannes

Les trois paramètres ajustés sont :

- ▶ le nombre de couches cachées
- ▶ le nombre de neurones par couche
- ▶ le taux d'apprentissage

La méthode de sélection est la suivante :

- ▶ on teste plusieurs niveaux de réduction en entrée (aucune, 50, 100, 125 composantes)
- ▶ on effectue une recherche par optimisation bayésienne pour chaque réduction

Recherche d'hyperparamètres sur Cannes

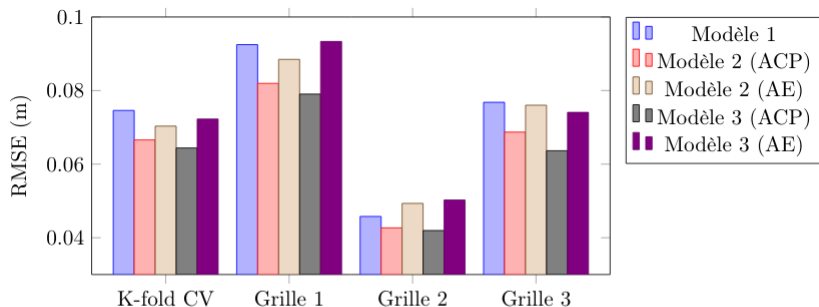
Les trois paramètres ajustés sont :

- ▶ le nombre de couches cachées
- ▶ le nombre de neurones par couche
- ▶ le taux d'apprentissage

La méthode de sélection est la suivante :

- ▶ on teste plusieurs niveaux de réduction en entrée (aucune, 50, 100, 125 composantes)
- ▶ on effectue une recherche par optimisation bayésienne pour chaque réduction
- ▶ on compare les performances des 3 meilleurs modèles de chaque recherche

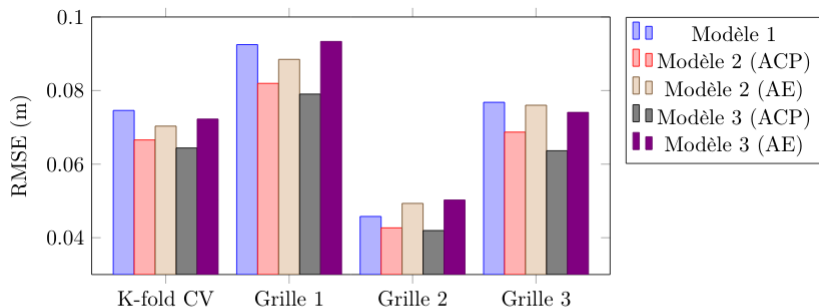
Recherche d'hyperparamètres sur Cannes



Modèle	Couches	Neurones	LR	Dimension d'entrée
Modèle 1	1	1024	$1 \cdot 10^{-4}$	Inchangée
Modèle 2	2	704	$4.64983 \cdot 10^{-4}$	50
Modèle 3	1	1024	$2.27403 \cdot 10^{-4}$	100

Figure: RMSE et paramètres des meilleurs modèles

Recherche d'hyperparamètres sur Cannes

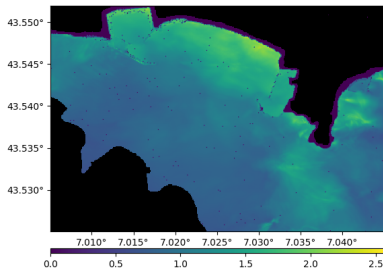


Modèle	Couches	Neurones	LR	Dimension d'entrée
Modèle 1	1	1024	$1 \cdot 10^{-4}$	Inchangée
Modèle 2	2	704	$4.64983 \cdot 10^{-4}$	50
Modèle 3	1	1024	$2.27403 \cdot 10^{-4}$	100

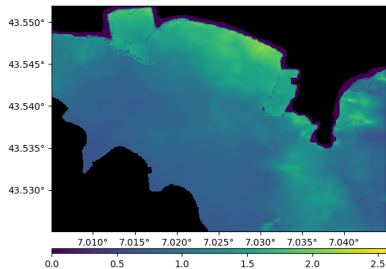
Figure: RMSE et paramètres des meilleurs modèles

→ un réseau à une couche avec une ACP en entrée donne les meilleurs résultats ici

Régularisation et overfitting



(a) Sans régularisation



(b) Dropout de 0.2

On observe un surapprentissage :

- ▶ application d'un dropout
- ▶ arrêt anticipé de l'entraînement
- ▶ la régularisation L^2 et la batch-normalization détériorent les résultats

Réseaux à couches de convolutions

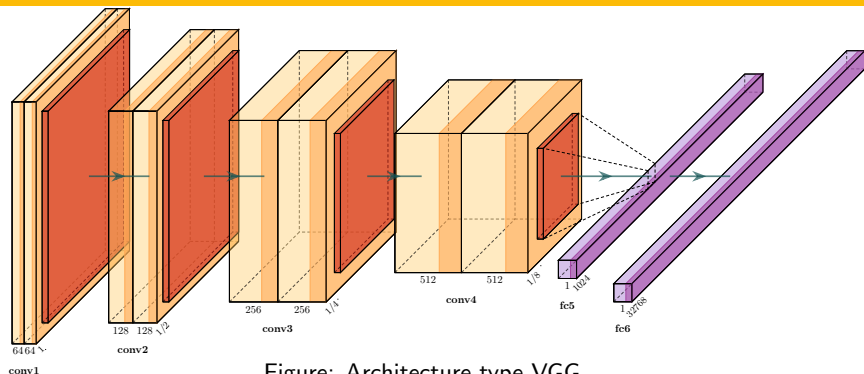


Figure: Architecture type VGG

Ajout de couches de convolution en entrée :

- ▶ augmentation du temps d'entraînement
- ▶ nouvelle recherche d'hyperparamètres peu concluantes

↔ réglage manuel des paramètres

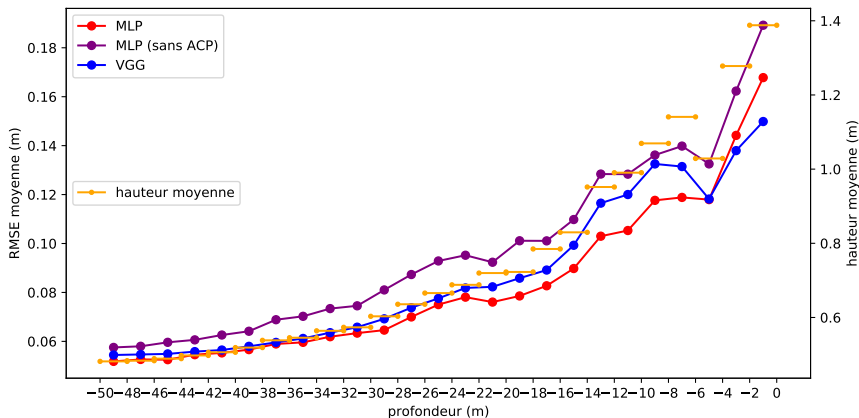


Figure: Erreur moyenne par tranche de profondeur, grille 5 de Nice (aéroport)

Globalement :

- ▶ assez bonne précision sous le niveau de la mer
- ▶ le MLP avec ACP reste le plus précis
- ▶ le VGG a tendance à présenter plus d'artefacts d'apprentissage

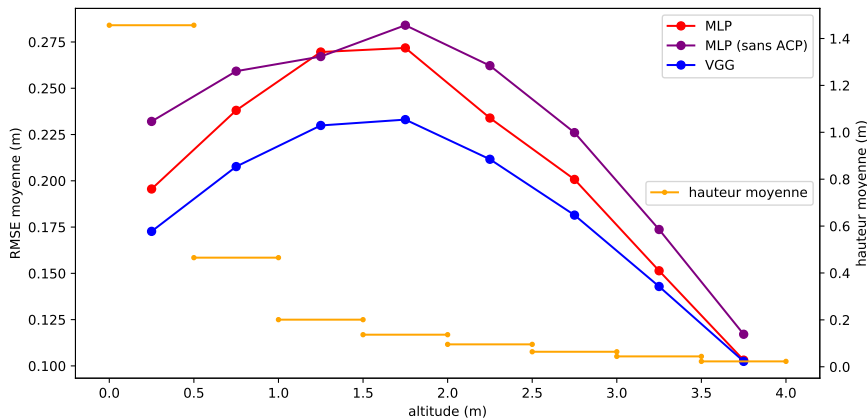
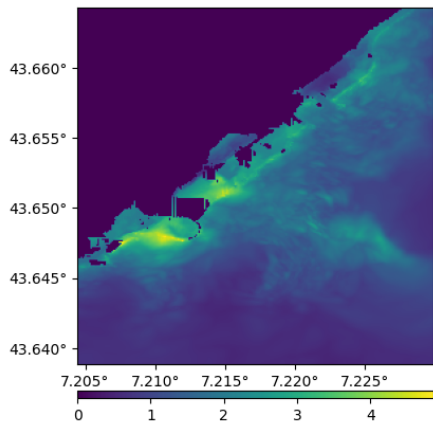


Figure: Erreur moyenne par tranche d'altitude, grille 5 de Nice (aéroport)

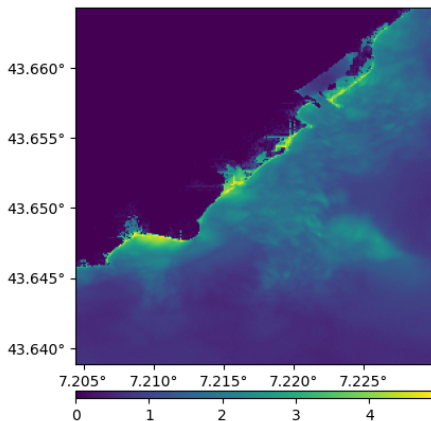
Au niveau des inondations :

- ▶ prédictions d'inondations très imprécises pour les fortes amplitudes
- ▶ le VGG présente des erreurs légèrement inférieures
- ▶ comportement similaire sur l'ensemble des sous grilles de sortie

Exemple de prédiction (MLP), magnitude 7.2 en zone 5



(a) Hauteurs cibles (m)



(b) Hauteurs prédites (m)

Sommaire

1. Introduction
2. Bases de données
3. Choix des architectures
4. Évaluation
 1. Sélection et évaluation d'un modèle
 2. Généralité du modèle
5. Conclusion

Sélection

Après comparaison des modèles MLP et VGG :

- ▶ le MLP avec ACP est plus performant en profondeur sur l'ensemble des grilles
- ▶ le VGG offre des résultats moins cohérents selon les grilles
- ▶ les prédictions d'inondation sont légèrement plus précises pour le VGG pour une poignée de scénarios

Sélection

Après comparaison des modèles MLP et VGG :

- ▶ le MLP avec ACP est plus performant en profondeur sur l'ensemble des grilles
- ▶ le VGG offre des résultats moins cohérents selon les grilles
- ▶ les prédictions d'inondation sont légèrement plus précises pour le VGG pour une poignée de scénarios

↪ on se concentre sur le MLP et on l'évalue sur :

- ▶ le jeu test de Nice
- ▶ le jeu de validation d'Antibes (non exploité)

Erreur selon la profondeur (hautes amplitudes)

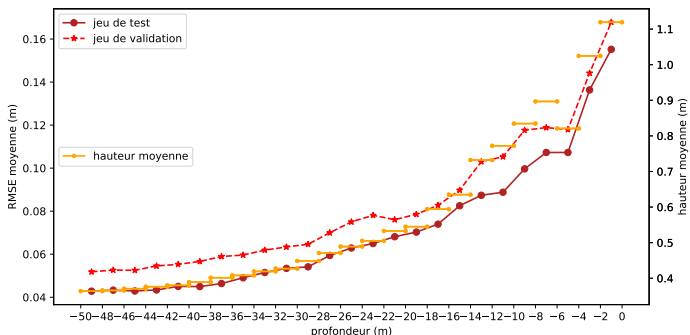
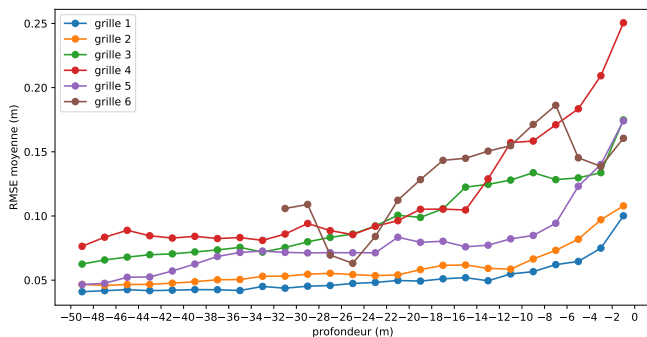


Figure: Erreurs moyennes sur la grille 5 (aéroport) de Nice

↪ même comportement sur les jeux de test et d'évaluation

Erreur selon la profondeur (hautes amplitudes)



(a) Erreurs moyennes sur le jeu de validation d'Antibes

Grille	1	2	3	4	5	6
Hauteur moyenne (m)	1.039	1.104	1.637	2.140	1.334	1.535

(b) Hauteurs simulées moyennes sur le jeu de validation d'Antibes

Erreur selon la hauteur de vague

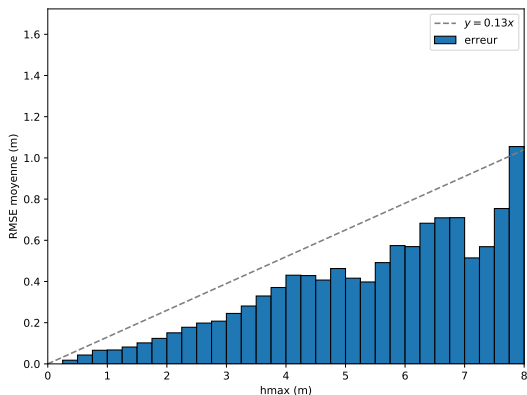
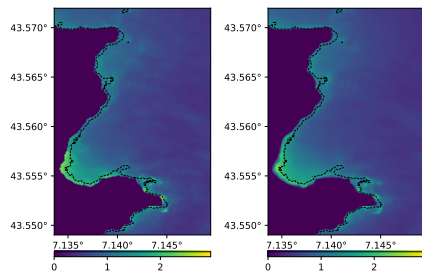


Figure: Erreur moyenne en fonction de hmax entre $-2m$ et $0m$ (hautes amplitudes, grille 4 d'Antibes)

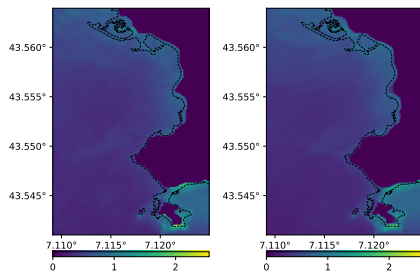
- ▶ hauteurs de vagues plus importantes sur Antibes
- ▶ l'erreur relative reste faible

Exemple de prédictions (Antibes)



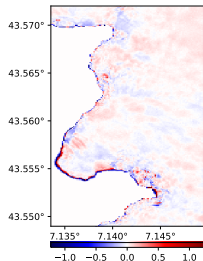
(a) Hmax cibles (m), G4

(b) Hmax prédits (m)

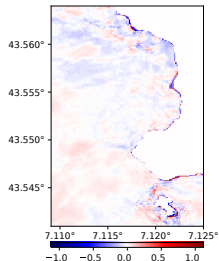


(c) Hmax cibles (m), G5

(d) Hmax prédits (m)



(e) Erreur signée, G4 (m)



(f) Erreur signée, G5 (m)

Généralité du modèle

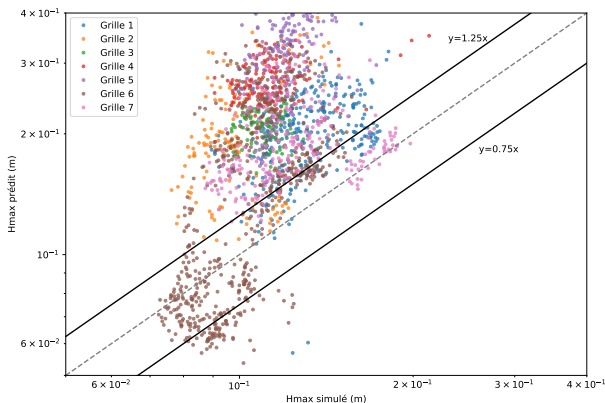


Figure: H_{max} prédites en fonction des hauteurs simulées sur Nice (magn. 8.5, zone de subduction hellénique)

- ▶ prédictions très imprécises pour un scénario sortant du domaine d'apprentissage
- ▶ faible capacité à généraliser

Sommaire

1. Introduction
2. Bases de données
3. Choix des architectures
4. Évaluation
5. Conclusion

Conclusion

Cette approche présente des avantages en contexte d'alerte :

- ▶ prédiction rapide en haute résolution (<1 min)
- ▶ bonne capacité à approcher les simulations en profondeur et près des côtes
- ▶ puissance de calcul nécessaire réduite pour les réseaux de neurones

Conclusion

Cette approche présente des avantages en contexte d'alerte :

- ▶ prédiction rapide en haute résolution (<1 min)
- ▶ bonne capacité à approcher les simulations en profondeur et près des côtes
- ▶ puissance de calcul nécessaire réduite pour les réseaux de neurones

Pendant, de nombreuses limites demeurent :

- ▶ forte dépendance aux données (qualité, quantité, variété des simulations)
- ▶ prédictions des fortes inondations peu précises
- ▶ comportement difficile à analyser
- ▶ modèles peu généralisables