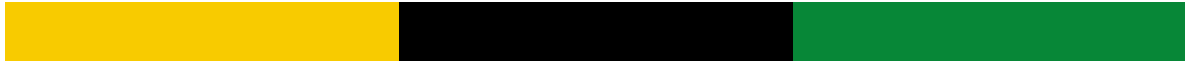




TÂCHE 4 : MACHINES À VECTEURS SUPPORTS POUR LA MODÉLISATION DE CRUES SUBITES

K. BOUKHAROUBA – P. ROUSSEL
G. DREYFUS

Laboratoire SIGMA, ESPCI ParisTech



OBJECTIFS

TÂCHE 4-1 :

“The goals of this sub-task are (i) to improve the modeling of the hydro-meteorological chain using SVMs, and (ii) to increase the knowledge and improve the design methodology of kernel-based dynamical models”.

Résultats : (i) modélisation de la relation pluie-hauteur, comparaison entre plusieurs méthodes de régression à noyaux ; (ii) nouvelle méthode de conception de modèles combinant apprentissage non supervisé (agrégation hiérarchique) et supervisé (SVM).

TÂCHE 4-2

“The issue that is addressed in this subtask is the integration of prior knowledge into dynamical support vector machines”.

Résultats : introduction de connaissances a priori par

- le prétraitement des données de pluie,
- la détermination des retards à prendre en considération pour la dynamique du modèle.

MODÈLES DYNAMIQUES

- Hypothèse NARX :

$$y(k+h) = f_h(x(k)) + d(k)$$

f_h est la fonction de régression (que l'on ne connaîtra jamais),
 d est une réalisation d'une variable aléatoire qui modélise les bruits et les perturbations,

x est le vecteur des variables du modèle, contenant des variables exogènes et des valeurs passées de y .

- Modèle postulé :

$$\hat{y}(k+h) = g_h(x(k), \theta)$$

g_h est le modèle postulé, muni des paramètres ajustables θ ,
 x est le vecteur des variables.

- Deux possibilités :

- ✓ Le vecteur $x(k)$ contient les valeurs passées des observations $y(k), \dots, y(k-n_a)$
 - ✓ Le vecteur $x(k)$ contient les valeurs passées des prévisions $\hat{y}(k), \dots, \hat{y}(k-n_a)$
- $x(k)$ contient en outre le vecteur des variables exogènes $u(k), \dots, u(k-n_b)$

MODÈLES POSTULÉS

$$g_n(\mathbf{x}(k), \theta)$$

- **MODÈLES LINÉAIRES** : linéaires par rapport à $\mathbf{x}$ et à θ .
- **MODÈLES POLYNOMIAUX** : non linéaires par rapport à $\mathbf{x}$, linéaires par rapport à θ .
- **MODÈLES NEURONAUX** : non linéaires par rapport à $\mathbf{x}$ et à θ .
- **MACHINES À VECTEURS SUPPORTS** : non linéaires par rapport à $\mathbf{x}$, linéaires par rapport à θ .

“ ε – INSENSITIVE SUPPORT VECTOR REGRESSION”

VERSION « HARD » :

Minimiser $1/2\|\theta\|^2$
sous la contrainte : **pour tous les exemples,**
l'erreur de modélisation est comprise entre $-\varepsilon$ et $+\varepsilon$

VERSION « SOFT » :

Minimiser $1/2\|\theta\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*)$
sous la contrainte : **pour toute exemple i**
l'erreur de modélisation est comprise entre $-\varepsilon - \xi_i^*$ et $\varepsilon + \xi_i$

C est la « constante de régularisation » (valeur déterminée par validation croisée).

LEAST SQUARES SVM

Régularisation « de Tikhonov » :

$$\text{Minimiser } \frac{1}{2}\|\theta\|^2 + C \sum_{i=1}^N (y_i - g(x_i, \theta))^2$$

sans contraintes.

Si le modèle postulé est linéaire en ses paramètres, il faut résoudre un système d'équations linéaires.

L'ASTUCE DES NOYAUX

Solution générale d'un problème de minimisation d'une fonction de coût

- calculée sur N exemples,
- comportant un terme de régularisation, pour un modèle postulé $g(\mathbf{x}, \theta)$:

$$g^*(\mathbf{x}, \theta) = \sum_{i=1}^N \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

On peut utiliser cette possibilité pour les deux méthodes et leurs variantes.

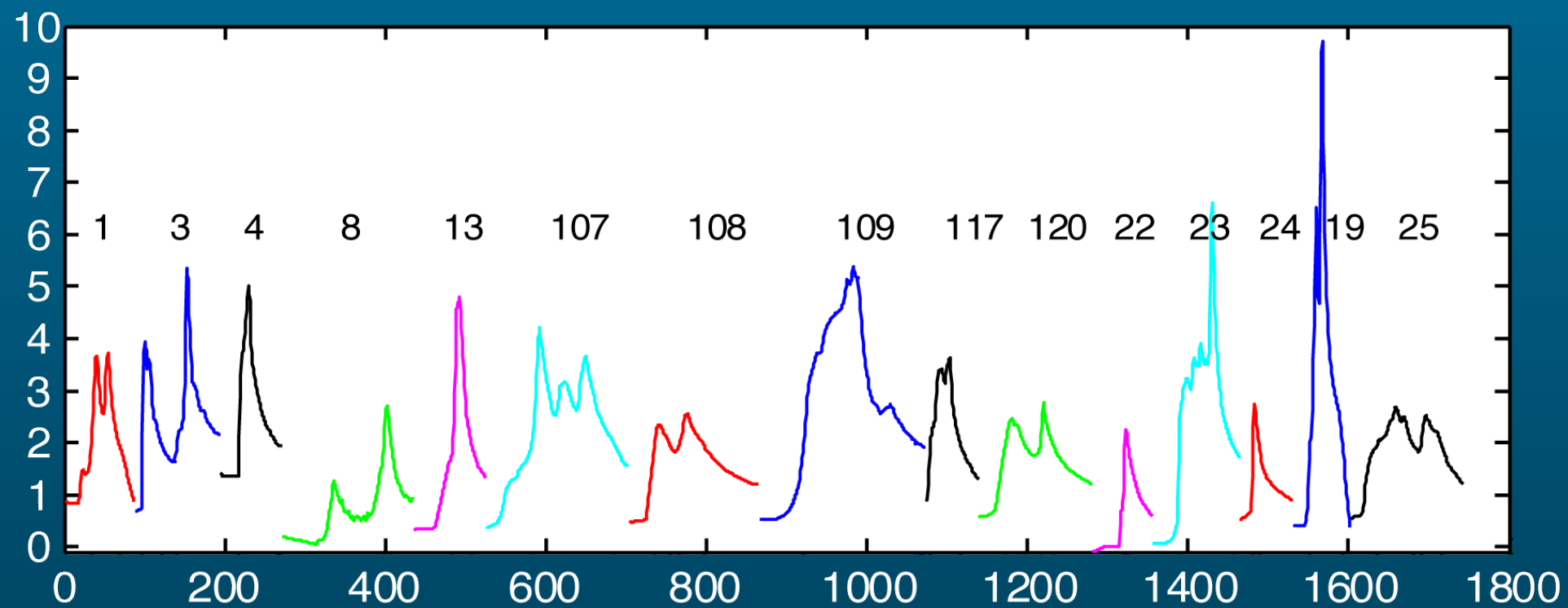
Noyau gaussien :

$$K(\mathbf{x}, \mathbf{x}_i) = \exp \left[-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2} \right]$$

AUTRES TYPES DE MODÈLES MIS EN ŒUVRE

- PieceWise Auto Regressive with eXogenous inputs (PWARX)
- Modèle de Hammerstein

BASE DE DONNÉES



CONCEPTION D'UN MODÈLE DE PRÉVISION GLOBAL

- **Apprentissage sur 12 événements**
- **Validation croisée partielle sur 4 autres événements (3, 4, 109, 23) pour estimer les hyperparamètres des modèles et, le cas échéant, sélectionner les variables (ordre du modèle et fenêtre de pluies).**
- **Test sur 4 événements (13, 19, 26, 27).**

Filtrage spatial des pluies

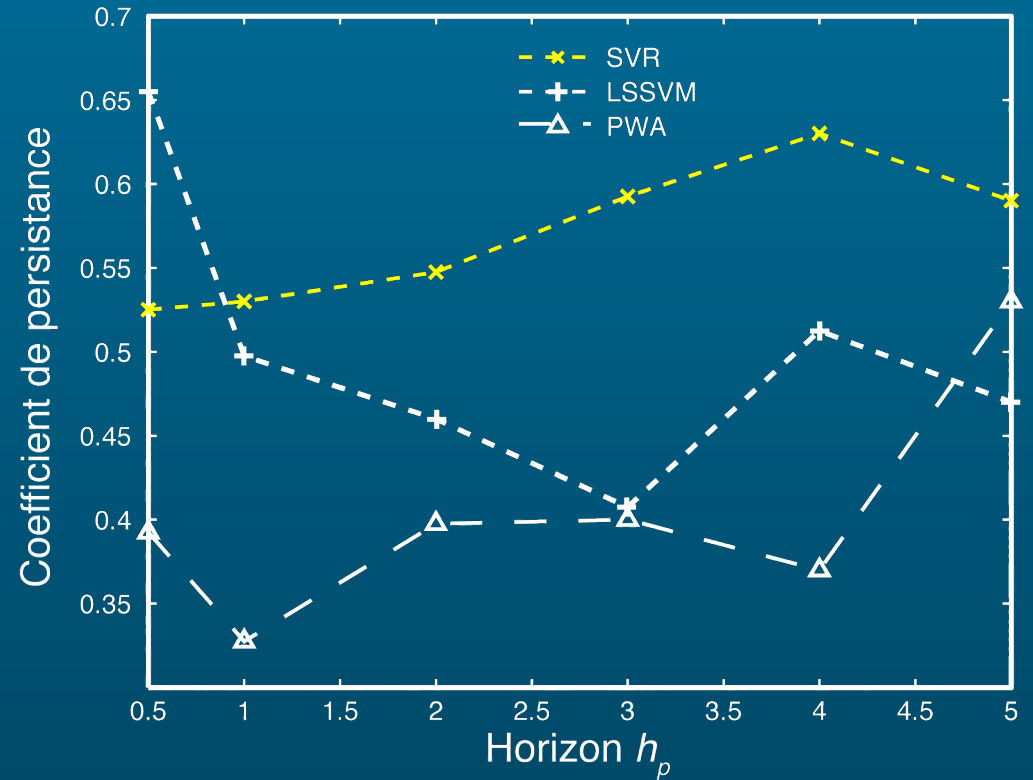
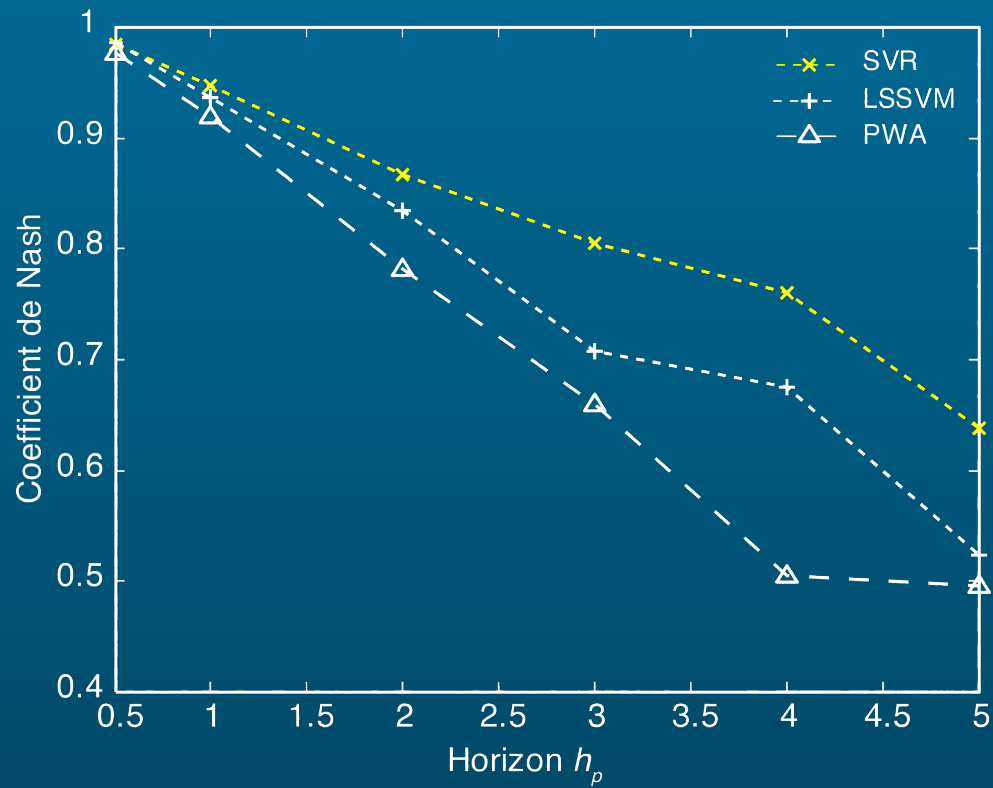
- *Sans filtrage spatial des pluies* : le vecteur des variables exogènes est la concaténation d'autant de vecteurs de pluies

$$\left[u_p(t), u_p(t-1), \dots, u_p(t-n_b) \right]$$

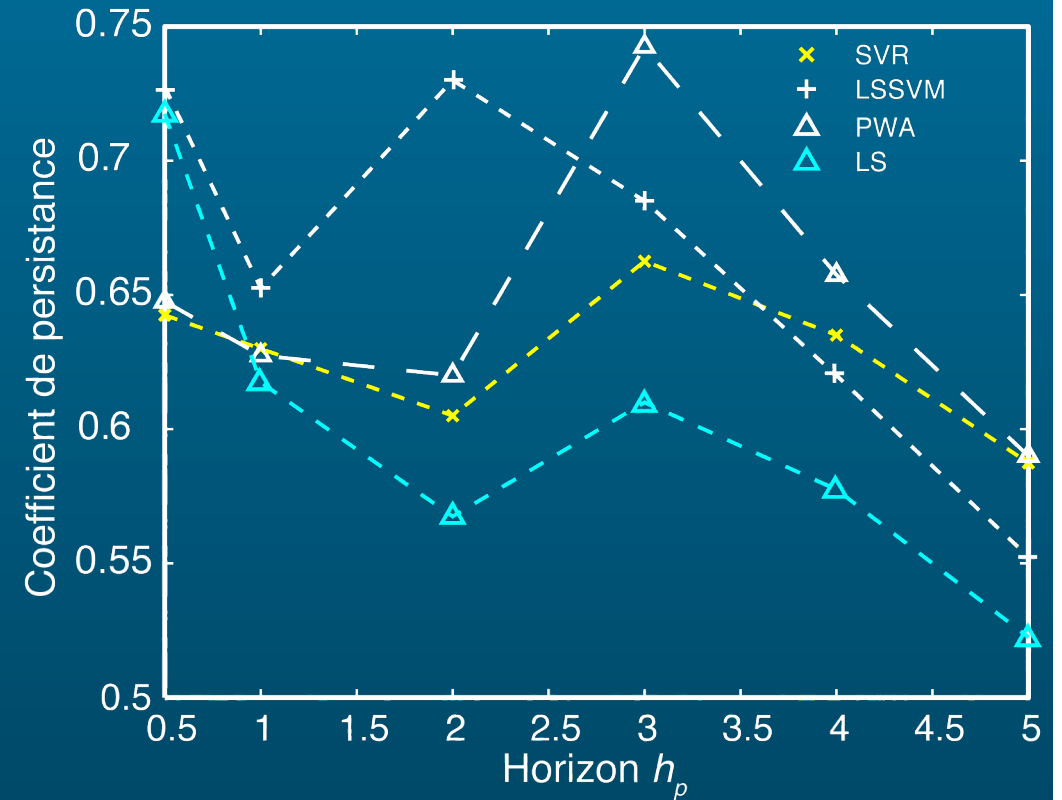
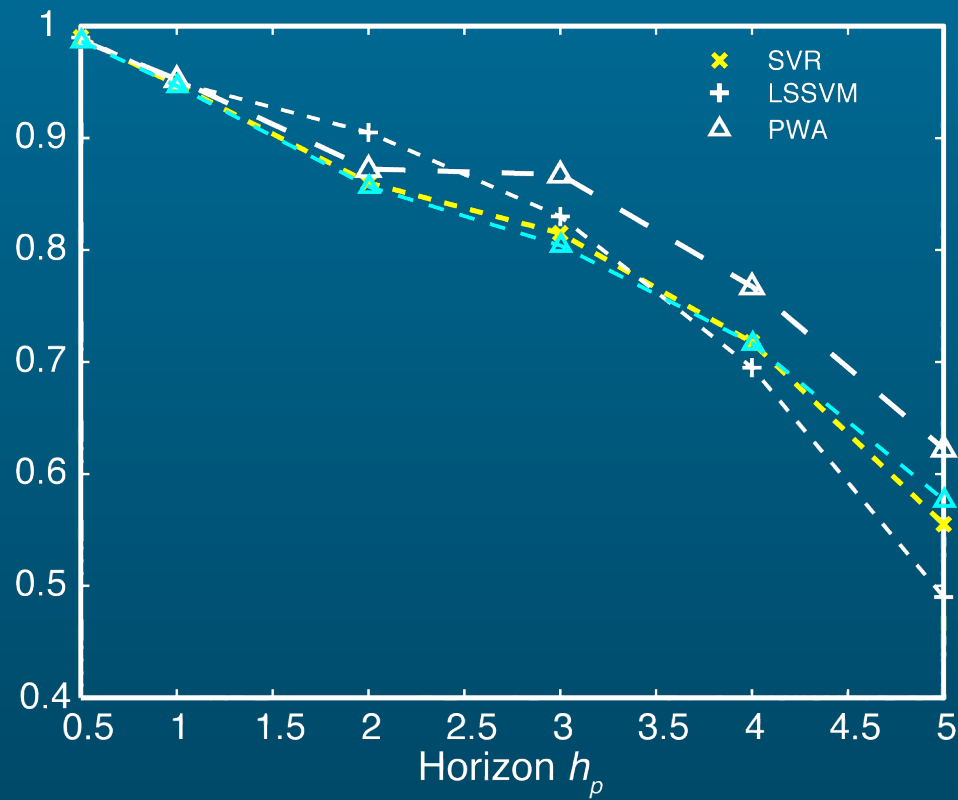
qu'il y a de pluviomètres.

- *Avec filtrage spatial des pluies* : un seul vecteur de hauteurs de pluie, calculé à l'aide des polygones de Thiessen.

MODÈLES GLOBAUX SANS FILTRAGE SPATIAL DES PLUIES



MODÈLES GLOBAUX AVEC FILTRAGE SPATIAL DES PLUIES

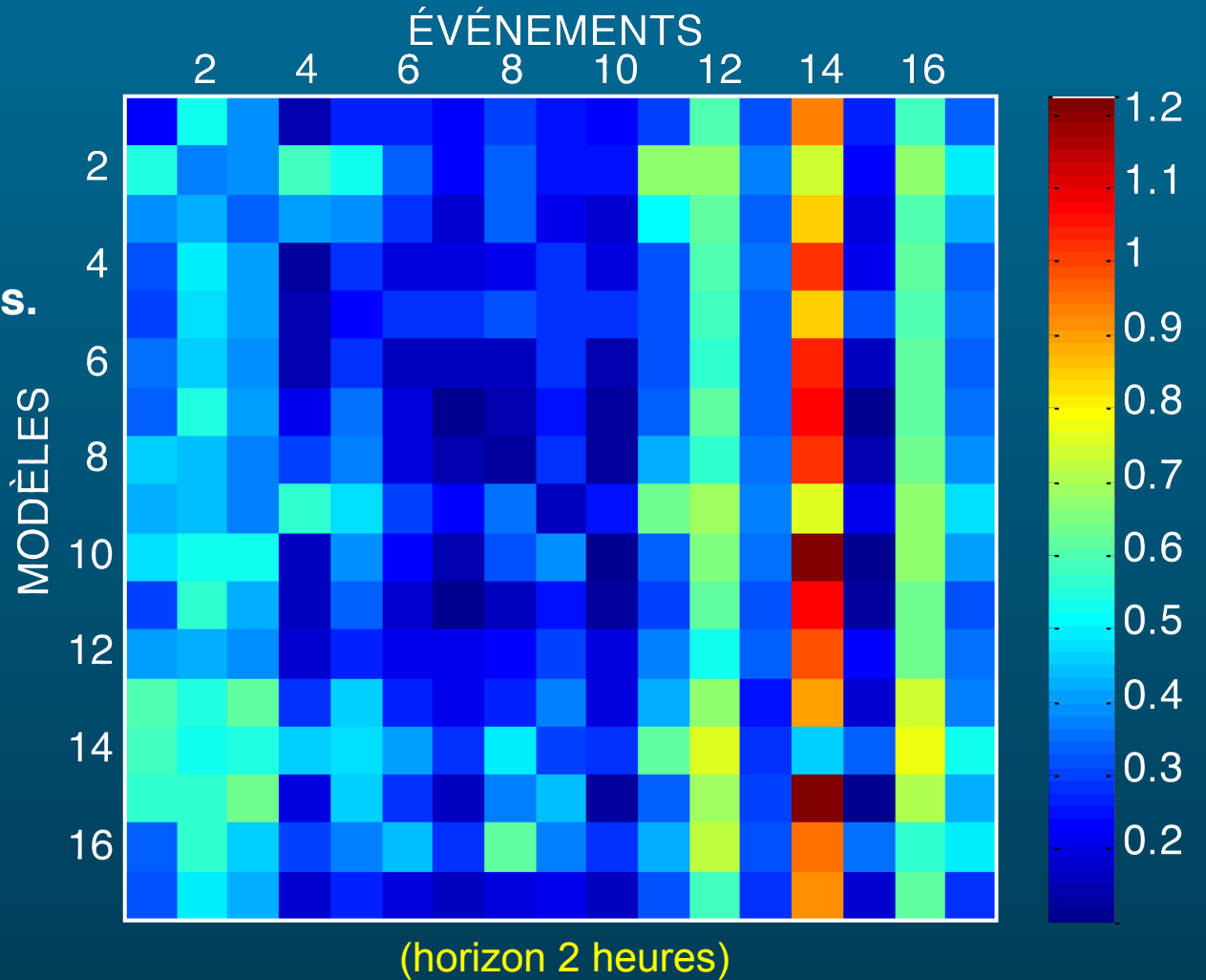


MODÈLES LOCAUX

PRINCIPE

- Construire un modèle par événement,
- agréger les événements dont les modèles ont des comportements analogues.

Distance modèle i - événement j : $d_{ij} = REQ M_{ij}$



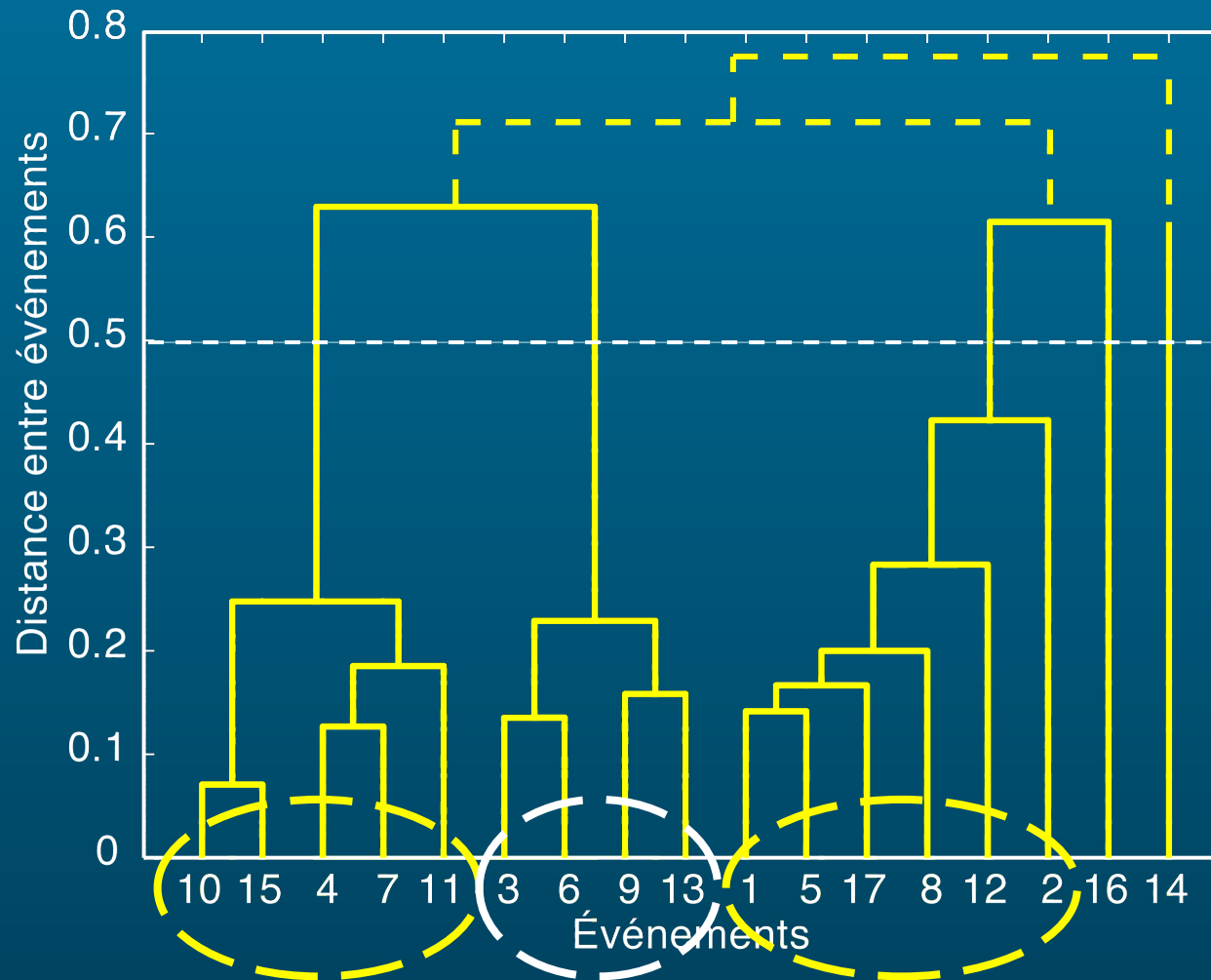
MODÈLES LOCAUX

AGRÉGATION DES MODÈLES

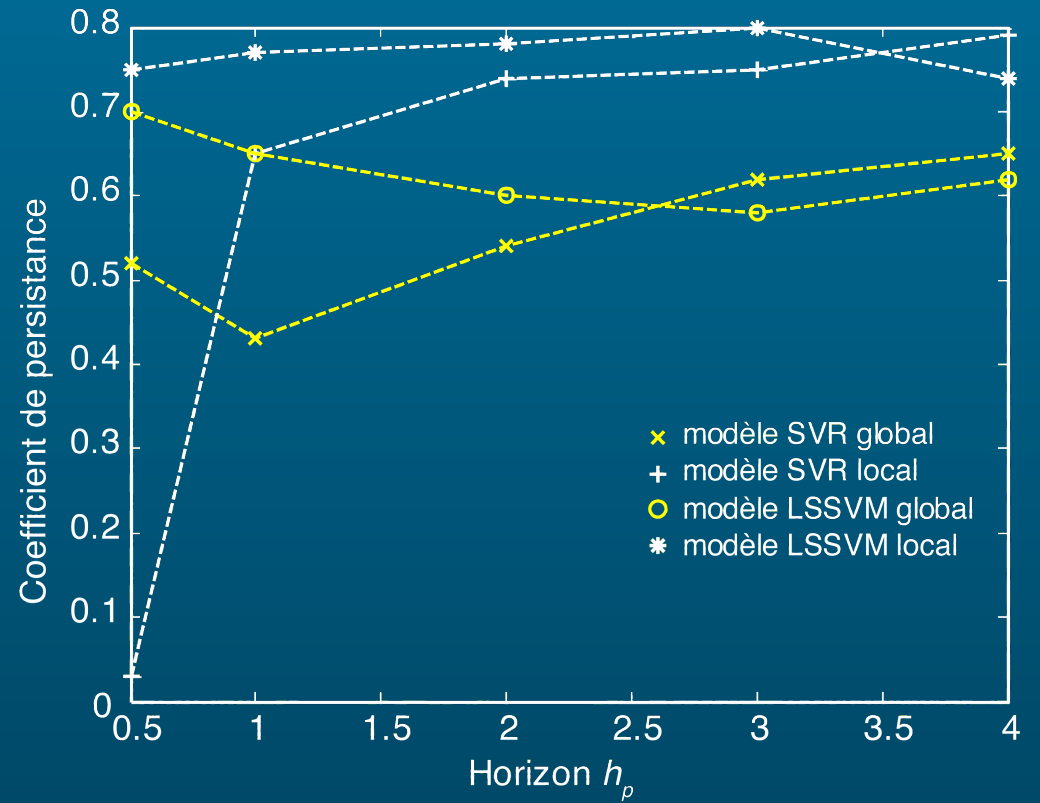
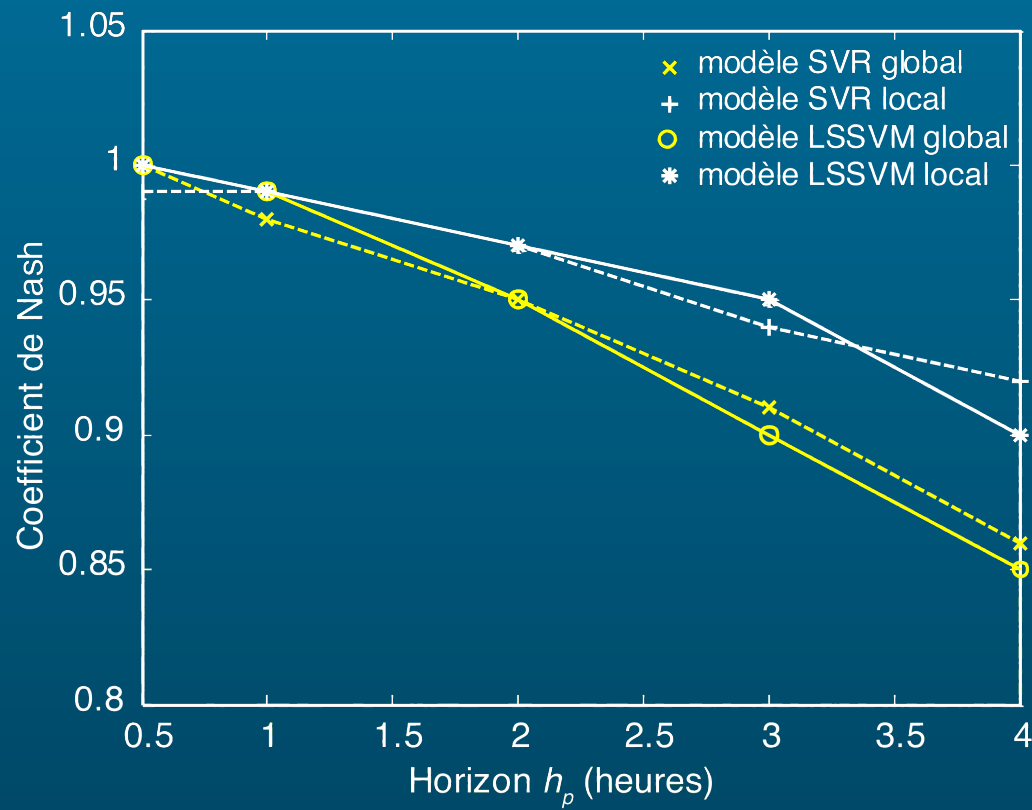
Distance événement i – événement $j = \max_{ij} (REQM_{ij}, REQM_{ji})$

Tant que tous les événements n'ont pas été agrégés
chercher dans la matrice des distances modèles-événements la paire d'événements
(k , l) les plus proches
construire un nouvel événement par concaténation des événements k et l
construire un modèle de ce nouvel événement
mettre à jour la matrice des distances modèles-événements
Fin

MODÈLES LOCAUX AGRÉGATION DES MODÈLES

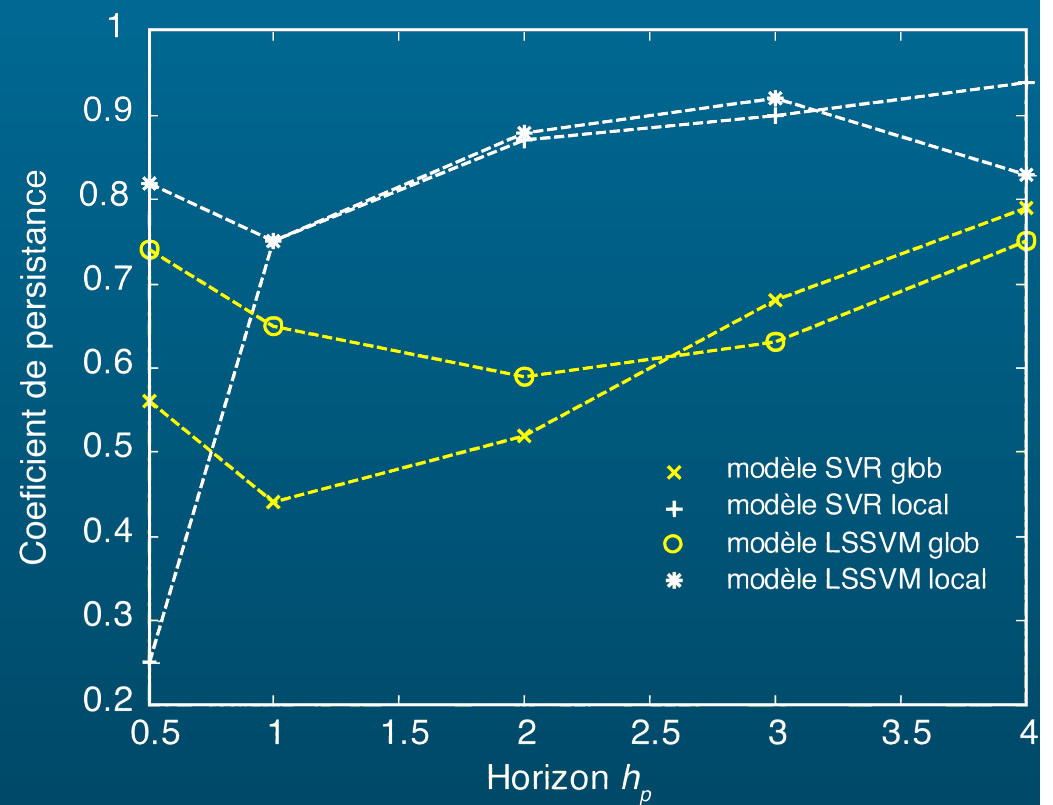
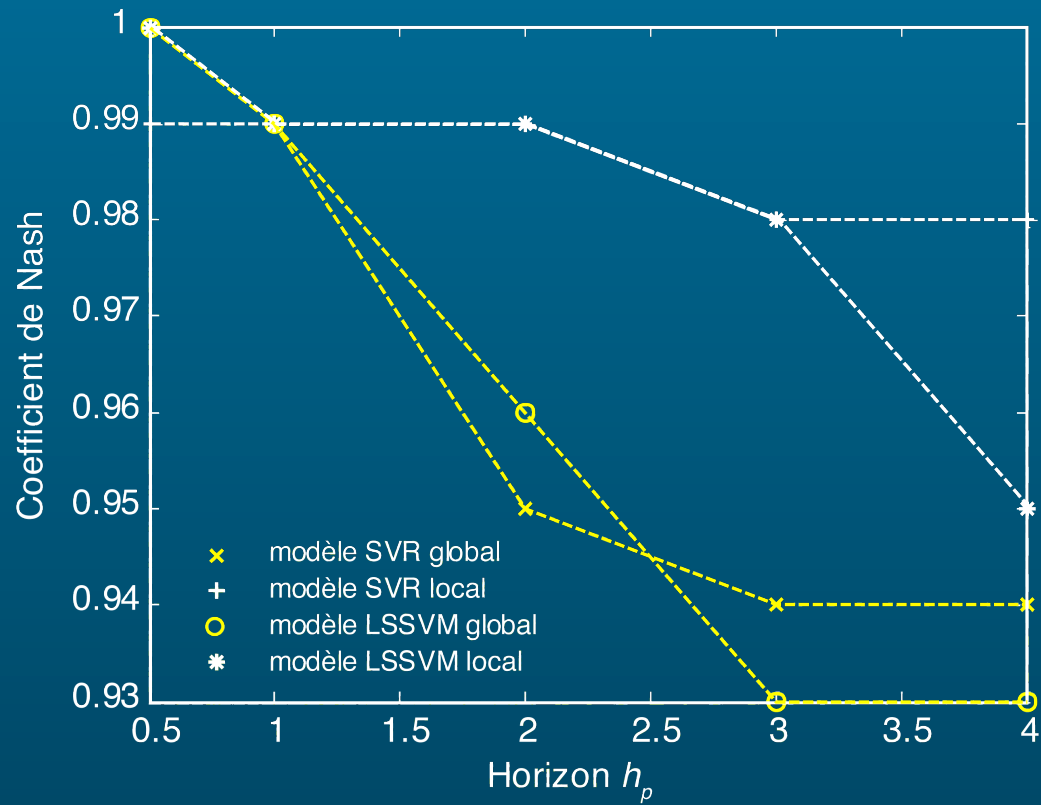


MODÈLES LOCAUX RÉSULTATS



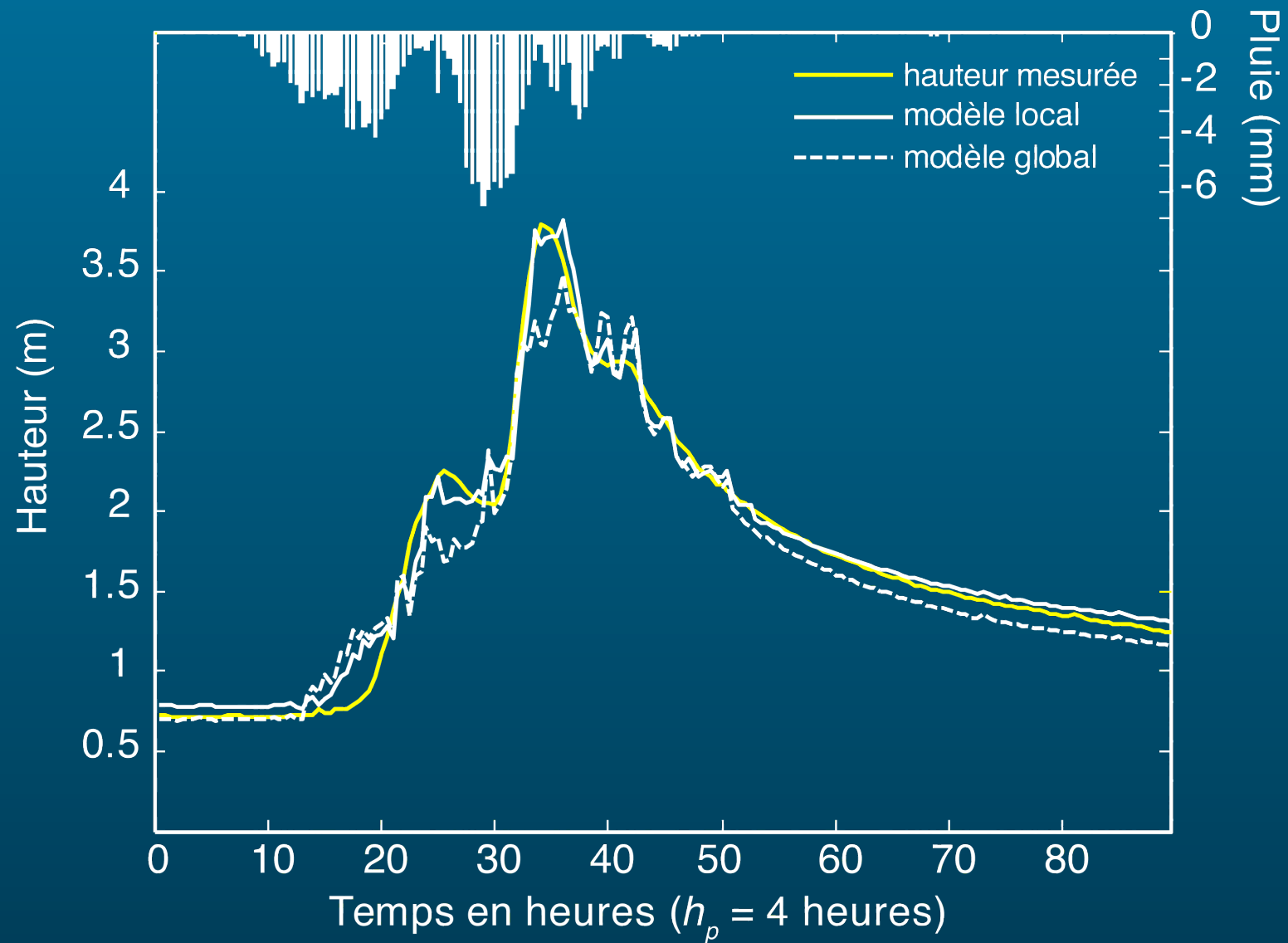
Événement 30

MODÈLES LOCAUX RÉSULTATS



Événement 32

MODÈLES LOCAUX RÉSULTATS



CONCLUSIONS ET PERSPECTIVES

CONCLUSIONS

- La conception de modèles locaux, par combinaison d'apprentissage non supervisé et d'apprentissage supervisé, permet une amélioration des performances par rapport à celles des modèles globaux.
- La prédiction de la hauteur d'eau et la position des pics sont très corrects.

PERSPECTIVES

- Approfondir l'interprétation des agrégats.
- Trouver une méthode efficace de sélection de modèle en cours d'événement.