



Exploration de l'assimilation de données multivariée pour la chimie atmosphérique

Antoine PEROT

Encadrants : Olivier PANNEKOUCKE et Vincent GUIDARD

Enseignant tuteur : François BACHOC

9 août 2021

Sujet :

À Météo-France, la prévision de la composition chimique de l'atmosphère est réalisée à l'aide du modèle de chimie transport MOCAGE (CTM) dont l'état initial est estimé à l'aide d'une chaîne d'assimilation de données propre au modèle. Actuellement, l'assimilation de données des espèces chimiques est construite en assimilant les espèces indépendamment les unes des autres, ce qui correspond à une assimilation univariée. Il s'agit là d'une approximation très grossière qu'on souhaite relaxer. Pour cela il faut envisager une assimilation multivariée, ce qui passe par la modélisation et l'estimation des covariances d'erreur de prévision inter-espèce (Emili et al. 2016). L'objectif du stage sera d'étudier différentes approches pour la modélisation et l'estimation des fonctions de covariance d'erreur de prévision multivariée, et la caractérisation de leur évolution temporelle. On s'appuiera pour cela sur le formalisme du filtre de Kalman paramétrique (Pannekoucke et al. , 2016 ; Pannekoucke et al. 2018) et sur un outil de génération automatique de réseau de neurone pour apprendre la dynamique à partir des données (Pannekoucke and Fablet, 2020 ; Pannekoucke, 2020 ; github repository).

Le stage se déroulera au CNRM et bénéficiera d'un environnement scientifique stimulant avec des interactions avec d'autres laboratoires (CERFACS, Institut Mines Télécom-Atlantique).

Remerciements :

PAS FAIT

J'aimerais remercier l'ensemble des enseignants qui ont été les miens à l'université Paul Sabatier, tout particulièrement lors de cette année 2019/2020 : messieurs Amodei, Malgouyres et Maréchal. Pour leur écoute et les connaissances qu'ils m'ont apportées je leur suis très reconnaissant.

Ensuite, je remercie tout particulièrement mes encadrants de stage Olivier Pannekoucke et Vincent Guidard qui m'ont guidé tout au long de ces six mois et m'ont apporté énormément de nouvelles connaissances etc...

Table des matières

1	Introduction aux problématiques de la chimie atmosphérique	3
2	Différentes méthode d'assimilation de données	3
2.1	L'assimilation de données dans la prévision de la composition chimique de l'atmosphère .	3
2.2	Filtre de Kalman et introduction aux méthodes d'ensembles	3
2.2.1	Le filtre de Kalman (KF)	3
2.2.2	Le filtre de Kalman d'ensemble (EnKF)	4
2.3	Illustration avec un modèle chimique jouet : Lotka-Volterra	4
2.3.1	Le modèle chimique de Lotka-Volterra	4
2.3.2	Exemple d'application de l'EnKF avec Lotka-Volterra	6
2.4	Une nouvelle méthode en assimilation de données : le filtre de Kalman paramétrique (PKF)	7
2.4.1	Principe	7
2.4.2	Illustration du PKF pour la prévision avec Lotka-Volterra	7
3	Transport 1D d'un polluant dans l'atmosphère	9
3.1	Formuler le PKF	9
3.2	Modèle de covariance univarié	10
3.3	Simulation numérique	10
3.3.1	Paramètres de simulation	10
3.3.2	Génération des champs d'erreurs aléatoires	11
3.3.3	Résultats : comparaison des prévisions du PKF et de l'EnKF	12
3.3.4	Résultats : comparaison des erreurs de prévisions pour le PKF et l'EnKF	13
3.4	Influence de l'erreur de modèle sur les diagnostics de l'EnKF	14
3.4.1	Mise en évidence d'un cas de divergence du filtre de Kalman d'ensemble	14
3.4.2	Impact de la réduction des longueurs de portées initiales sur les diagnostics . .	14
3.5	Conclusion dans le cadre univarié de l'advection 1D	15
4	Un modèle de chimie transport simplifié pour l'exploration d'une formulation multivariée du PKF	16
4.1	Formulation multivariée du PKF avec le SymPKF	16
4.2	Simplification du système PKF	18
4.3	Simulation numérique : une prévision multivariée du système PKF	19
4.3.1	Paramètres de simulation	19
4.3.2	Résultats : comparaison des prévisions PKF et EnKF	19
4.3.3	Résultats : comparaison des erreurs de prévisions PKF et EnKF	19
4.4	Définition d'un modèle de corrélation multivarié	20
4.4.1	Diagnostics à l'aide d'un ensemble	20
4.4.2	Une première formulation analytique	21
4.4.3	Une voie d'amélioration du modèle de corrélations croisées	23
4.4.4	Un modèle de corrélations croisées hybride : analytique et réseau de neurones . .	24
4.5	Assimilation d'une observation : PKF et EnKF	27
5	Vers un modèle chimique transport plus réaliste	27
5.1	Introduction du modèle chimique GRS	27
5.2	Comparaison prévision PKF et EnKF	28
	Références	31

1 Introduction aux problématiques de la chimie atmosphérique

bla bla bla

2 Différentes méthode d'assimilation de données

2.1 L'assimilation de données dans la prévision de la composition chimique de l'atmosphère

L'assimilation de données est au coeur des problématiques de chimie atmosphérique, ou même en prévision numérique du temps (PNT). Il s'agit, partant à $t = 0$ d'une connaissance a priori de l'état de l'atmosphère, notée $x_0 = x^b$ (b pour *background*), aussi appelée "ébauche" et d'une erreur d'ébauche $\Sigma_0 = \Sigma^b$, d'estimer les états futurs de l'atmosphère $(x_t)_{t>0}$ à l'aide d'un modèle de prévision \mathcal{M} et leurs erreurs associées $(\Sigma_t)_{t>0}$. A intervalles réguliers (toutes les 6h), il faut mettre à jour la prévision à l'aide des observations recueillies entre temps. Ces observations peuvent provenir de capteurs sur terre, en mer, dans les airs (ballons ou avions de ligne) ou encore de satellites. L'étape de prise en compte des observations est appelée "phase d'analyse". Ainsi, l'assimilation de données consiste en une suite de cycles de prévision/analyse.

Le filtre de Kalman [2] (KF) est un algorithme à la base de l'assimilation de données bien que pas utilisé dans les problématiques étudiées ici, pour des raisons qui seront explicitées plus tard. Il existe principalement deux alternatives aux KF datant des années 1990 et qui sont utilisées aujourd'hui dans le monde. Dans l'ordre chronologique, les méthodes d'ensemble (EnKF) [?] puis les méthodes variationnelles (paper 3DVar, 4DVar). De plus en plus, des hybridations de ces deux méthodes sont formulées (citer un truc...) afin de tirer le meilleur de chacune des deux méthodes. Nous nous intéresserons ici principalement aux méthodes d'ensemble.

Plus récemment [5] a été formulée une nouvelle méthode d'assimilation de données, le filtre de Kalman paramétrique (PKF) qui sera l'objet principal de ce stage et dont nous détaillerons plus tard le principe, les apports et les difficultés qui lui sont propre.

2.2 Filtre de Kalman et introduction aux méthodes d'ensembles

Nous introduisons ou ré-introduisons en suivant les principes du filtre de Kalman, afin d'ensuite présenter le filtre de Kalman d'ensemble (EnKF), ses avantages et ses limites.

2.2.1 Le filtre de Kalman (KF)

$$\begin{cases} \mathbf{y}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t, & \mathcal{N}_{m_t}(0, \mathbf{R}_t) \\ \mathbf{x}_t = \mathbf{M}_t \mathbf{x}_{t-1} + \mathbf{w}_t, & \mathcal{N}_n(0, \mathbf{Q}_t) \\ \mathbf{x}_0 = \mathbf{x}^b, \Sigma^a = \Sigma^b \end{cases}$$

où \mathbf{y}_t est le vecteur m_t dimensionnel d'observations au temps t , \mathbf{H}_t l'opérateur d'observation, \mathbf{M}_t le modèle de prévision au temps, \mathbf{x}_t le vecteur n dimensionnel d'un état qu'on cherche à estimer. On fait l'hypothèse que les matrices $\mathbf{R}_t, \mathbf{Q}_t$ sont connues. Tous les opérateurs (modèle, observation, erreur de modèle, erreur d'observation) peuvent dépendre du temps, et sont donc indexés par t .

Un cycle d'assimilation, qui consiste en deux étapes, une de prévision (f pour *forecast*), puis une phase d'analyse (a pour *analysis*), suit la dynamique suivante dans le filtre de Kalman :

$$\text{KF - forecast : } \begin{cases} \mathbf{x}_t^f = \mathbf{M}_t \mathbf{x}_{t-1}^a & (1a) \\ \Sigma^f = \mathbf{M}_t \Sigma_{t-1}^a \mathbf{M}_t^T + \mathbf{Q}_t & (1b) \end{cases}$$

$$\text{KF - analysis : } \begin{cases} \mathbf{K}_t = \Sigma_t^f \mathbf{H}_t^T (\mathbf{H}_t^T \Sigma_t^f \mathbf{H}_t + \mathbf{R}_t)^{-1} & (2a) \\ \mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t (\mathbf{y}_t - \mathbf{H}_t \mathbf{x}_t^f) & (2b) \\ \Sigma_t^a = (\mathbf{I}_n - \mathbf{K}_t \mathbf{H}_t) \Sigma_t^f & (2c) \end{cases}$$

En résumé, \mathbf{x}_t^f est la prévision fournie par le modèle \mathbf{M}_t , \mathbf{x}_t^a est la correction de celle-ci obtenue après prise en compte judicieuse grâce au gain de Kalman \mathbf{K}_t des observations \mathbf{y}_t . \mathbf{K}_t le gain de Kalman détermine le niveau de prise en compte des observations en faisant la balance entre erreurs de prévision, erreur de modèle et erreur d'observation. La matrice de covariance d'erreur de prévision Σ_t^f est finalement mise à jour en fonction de la matrice d'erreur d'observation (ligne 1b).

2.2.2 Le filtre de Kalman d'ensemble (EnKF)

Dans les modèles géophysiques, l'opérateur \mathbf{M} n'est en général pas linéaire. Comme les modèles de prévisions sont de la forme :

$$\partial_t \mathbf{X} = \mathcal{M}(t, \mathbf{X})$$

et sont discrétisés en temps et en espace ; on pourrait approximer \mathbf{M}_t dans les équations du filtre de Kalman par la multiplication successive des opérateurs tangents linéaires (dérivées directionnelles) de \mathcal{M} en chaque pas de temps de $t-1$ à t , mais cela serait trop coûteux en raison de la trop grande dimension des problèmes (10^6 à 10^7).

Le filtre de Kalman d'ensemble est alors une extension non linéaire du filtre de Kalman qui consiste en une approximation de la matrice de covariance d'erreur de prévision Σ_t^f par la matrice de covariance empirique d'un échantillon, appelé "ensemble", de prévisions $(\tilde{\mathbf{x}}_t^i)_{i=1,\dots,N}$ suivant la dynamique :

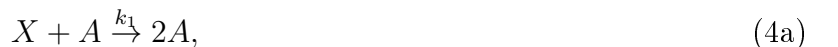
$$(EnKF) : \begin{cases} \mathbf{x}_0^i \sim \mathcal{N}(\mathbf{x}^b, \Sigma^b) & (3a) \\ \mathbf{x}_t^{f,i} = \mathcal{M}_{t-1 \rightarrow t}(\mathbf{x}_{t-1}^{a,i}) + \mathbf{w}_t^i & (3b) \\ \mathbf{x}_t^{a,i} = \mathbf{x}_t^{f,i} + \mathbf{K}_t (\mathbf{y}_t + \mathbf{v}_t^i - \mathbf{H}_t \mathbf{x}_t^{f,i}) & (3c) \end{cases}$$

Notons que cette méthode d'assimilation présente des avantages importants : elle permet de considérer des modèles non linéaires sans difficulté, elle est très simple à mettre en oeuvre puisqu'il ne suffit de faire que plusieurs fois une même prévision, de ce fait elle est parallélisable, ce qui permet une réduction considérable des temps de calcul.

2.3 Illustration avec un modèle chimique jouet : Lotka-Volterra

2.3.1 Le modèle chimique de Lotka-Volterra

On considère s'intéresse à quatre espèces chimiques A, B, X et Y soumises aux réactions chimiques suivantes :



Cette suite de réaction revient à la réaction $X \rightarrow Y$. La dynamique des concentrations de A et B peut-être modélisée par le système d'équations différentielles suivant :

$$\frac{d[A]}{dt} = k_1[X][A] - k_2[A][B] \quad (5a)$$

$$\frac{d[B]}{dt} = k_2[A][B] - k_3[B] \quad (5b)$$

où $[A]$ désigne la concentration de l'espèce chimique A . Si la concentration dans l'atmosphère de l'espèce chimique X est constante (c'est à dire en excès), le système se réécrit :

$$(LV) : \begin{cases} \frac{d[A]}{dt} = k_1[A] - k_2[A][B] \\ \frac{d[B]}{dt} = k_2[A][B] - k_3[B] \end{cases} \quad (6a)$$

Les paramètres k_1, k_2 et k_3 correspondent respectivement aux vitesses de reproduction de A , de réaction pour les réactifs A et B , et de désintégration de B en Y . On considèrera arbitrairement que $k_1 = 0.075, k_2 = 0.065$ et $k_3 = 0.085$ pour notre modèle \mathcal{M} .

Méthode de résolution numérique : afin de simuler le système (6) qui est un problème de Cauchy, on développe une classe en Python capable de simuler tous les problèmes de cette forme. On résout numériquement ces équations en appliquant au choix les méthodes d'Euler Explicite, Runge-Kutta 2 ou 4.

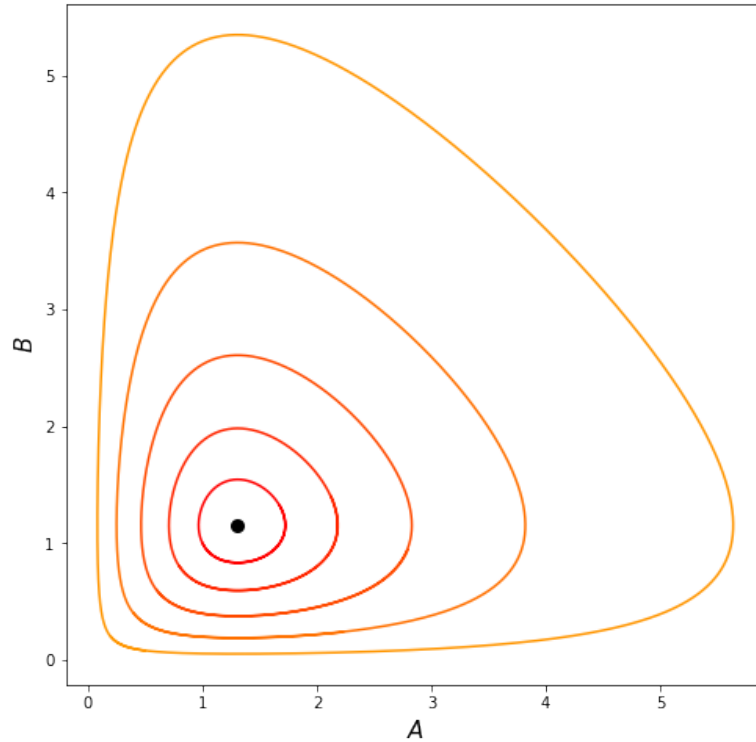


FIGURE 1 – Exemples de simulations

Les trajectoires dans le plan de phase de Lotka-Volterra correspondent à des orbites non circulaires dans le cadran $\{x > 0, y > 0\}$ et autour du point stationnaire $\left(\frac{k_3}{k_2}, \frac{k_1}{k_2}\right)$.

2.3.2 Exemple d'application de l'EnKF avec Lotka-Volterra

A présent, on souhaite simuler des cycles de prévision/assimilation à l'aide de l'EnKF. Notre modèle de prévision \mathcal{M} reste inchangé, mais l'on considère par soucis de réalisme qu'il ne représente pas les vraies lois de la nature \mathcal{N} . On considère par soucis de simplicité que \mathcal{N} est un modèle de type Lotka-Volterra dont les paramètres diffèrent légèrement de ceux de notre modèle \mathcal{M} .

On simule alors quelques cycles de prévision/assimilation. Pour cela, on considère un ensemble de 15 membres, un état initial \mathbf{x}^b et matrice de covariance initiale Σ^b tous deux connus. A intervalles réguliers, on recueille une observation de la réalité dont on connaît le bruit. Finalement, on applique les équations (3).

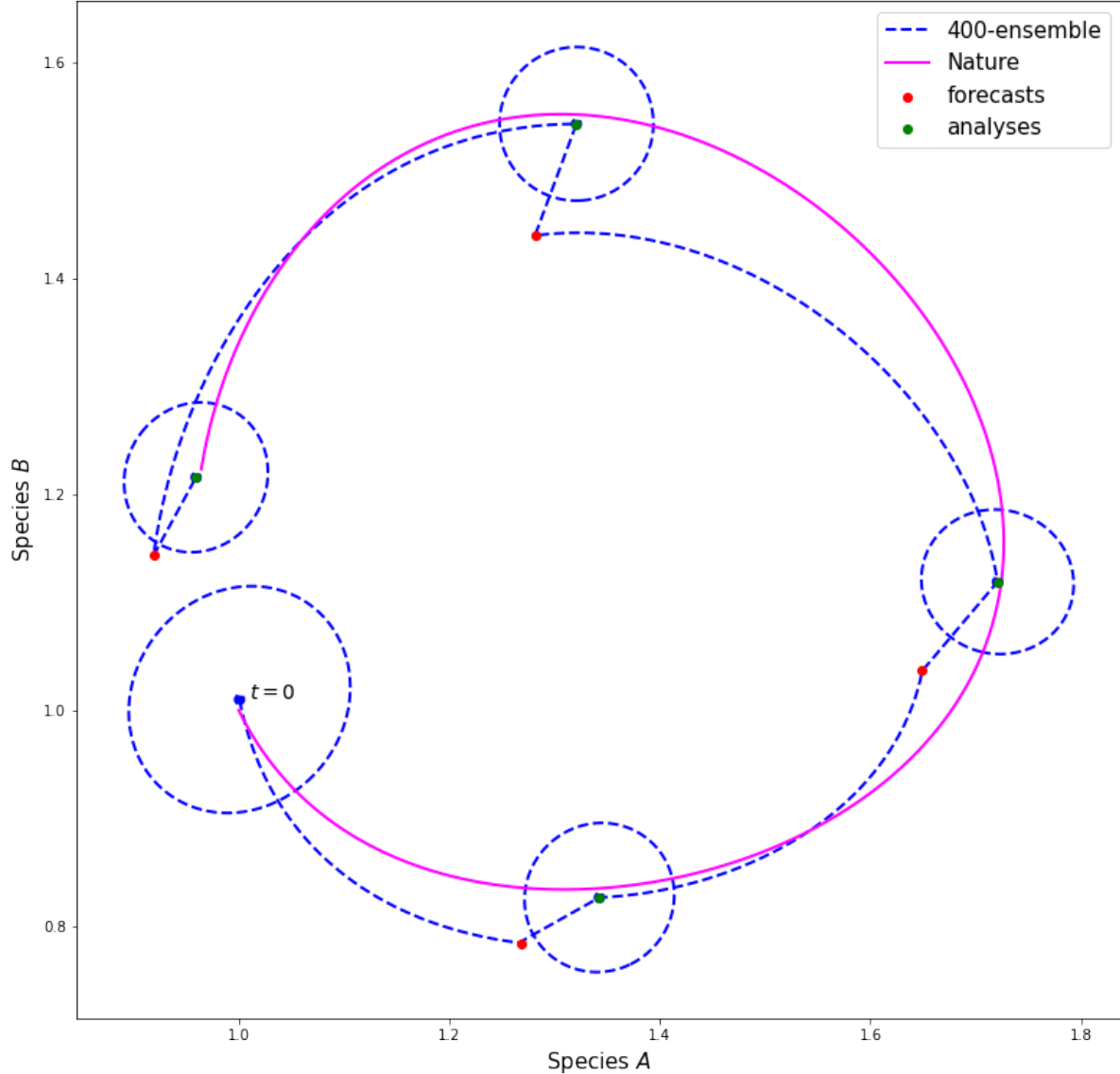


FIGURE 2 – Simulation de 4 cycles d'assimilation avec l'EnKF

Les étapes d'assimilation des observations nous permettent de rester sur une trajectoire raisonnablement proche de la réalité.

2.4 Une nouvelle méthode en assimilation de données : le filtre de Kalman paramétrique (PKF)

2.4.1 Principe

Après avoir évoqué (section 2.2.2) la nécessité de considérer les méthodes d'ensembles comme alternative au filtre de Kalman classique, il convient de dire que le filtre de Kalman d'ensemble souffre à de certaines limites.

De part sa nature, il est sujet à un bruit d'échantillonnage d'autant plus important que les ensembles considérés sont de petites tailles (de la trentaine à la centaine de membre dans les cadres opérationnels, petits donc). Ce bruit d'échantillonnage vient donc polluer les diagnostics des prévisions (moyenne empirique des prévisions) et des matrices de covariances empiriques d'erreurs de prévision. Pour pallier à ces défauts, des méthodes de localisation ont été mises au point afin d'éviter que le bruit d'échantillonnage ne viennent indiquer des corrélations entre des paires de points trop distants sur la carte. A l'inverse, les diagnostics effectués sur l'ensemble peuvent être grandement erronés par l'erreur de modèle, en pratique inconnue, comme nous le verrons dans la section 3.4. Enfin, en raison de la trop grande dimension des problèmes, il est impossible de manipuler directement les matrices de covariance. Estimer correctement l'erreur de prévision est pourtant crucial afin de fournir des prévisions de la meilleure qualité possible.

Le PKF [5] est une nouvelle méthode d'assimilation qui vient décrire la dynamique des incertitudes d'une manière novatrice : les matrices de covariance y sont approchées par des modèles de covariance reposant sur un jeu léger de paramètres. Le filtre de Kalman paramétrique consiste alors en la description de la dynamique de ces paramètres afin d'ensuite estimer l'estimer les fonctions de covariance et permettre l'assimilation des observations. On illustre en suivant le PKF dans un cadre adimensionnel et le compare à l'EnKF, afin de mieux en saisir les apports.

2.4.2 Illustration du PKF pour la prévision avec Lotka-Volterra

Dans un cadre bivarié (deux espèces chimiques) et adimensionnel, on illustre le PKF. Dans un tel contexte, l'état que l'on cherche à prévoir est de la forme (A^f, B^f) , et sa matrice de covariance Σ^f : s'écrit :

$$\Sigma^f = \begin{pmatrix} V_A^f & V_{AB}^f \\ V_{AB}^f & V_B^f \end{pmatrix} \quad (7)$$

Ainsi, l'incertitude associée la prévision est caractérisée par les 3 quantités V_A^f, V_B^f et V_{AB}^f . Afin d'estimer la dynamique de ces paramètres, on procède par décomposition de Reynolds. On décompose un état initial (A, B) comme étant un état moyen perturbé, c'est à dire : $(A, B) = (\bar{A}, \bar{B}) + (\varepsilon_A, \varepsilon_B)$ où (\cdot) désigne la moyenne, l'espérance. On peut ensuite définir les paramètres d'incertitude de Σ^f par $(V_A, V_B, V_{AB}) = (\overline{\varepsilon_A^2}, \overline{\varepsilon_B^2}, \overline{\varepsilon_A \varepsilon_B})$.

Ensuite, en utilisant les équations du problème (6) et des manipulations algébriques, on parvient au système suivant :

$$(PKF - LV) : \begin{cases} \partial_t A = aA - bAB - bV_{AB} & (8a) \\ \partial_t B = -cB + dAB + dV_{AB} & (8b) \\ \partial_t V_A = 2(aV_A - bV_{AB}A - bBV_A + \overline{b\varepsilon_A^2\varepsilon_B}) & (8c) \\ \partial_t V_B = 2(-cV_B + dV_{AB}B + dbAV_B + \overline{b\varepsilon_B^2\varepsilon_A}) & (8d) \\ \partial_t V_{AB} = V_{AB}(a - bB - c + dA) - bV_BA + dV_AB + \overline{d\varepsilon_A^2\varepsilon_B} - \overline{b\varepsilon_A\varepsilon_B^2} & (8e) \end{cases}$$

Notez que dans toutes les équations (8) on décrit la dynamique des états moyens (\bar{A}, \bar{B}) bien que l'on ait omis le symbole de l'opérateur d'espérance : par souci de lisibilité, il s'agira toujours de l'état moyen

dans les systèmes du PKF. Ensuite, les termes en rouge correspondent à des termes non fermés d'ordre 3. Le PKF se limite à une description des moments d'ordre au plus 2 (tangent-linéaire) des paramètres d'incertitudes, ainsi ces termes seront négligés.

Enfin, le calcul à la main de la dynamique des paramètres du PKF peut vite s'avérer fastidieux et source d'erreur, en particulier dès que l'on considère des problèmes d'une plus grande complexité. Pour cette raison a été développé un package en langage Python, nommé SymPKF [4], dont nous nous servons dans la suite de ce rapport, et qui se base sur du calcul symbolique pour déterminer de manière automatisée la dynamique du système du PKF pour un problème donné.

On compare à présent les prévisions fournies par le PKF avec celle fournies par un ensemble de taille 1600 (erreur d'échantillonnage 2.5%). Initialement, on considère que $\mathbf{x}_0 = (1, 1)$ et $\Sigma_0 = 0.01\mathbf{I}_2$ et l'on génère l'ensemble initial selon la loi $\mathcal{N}(\mathbf{x}_0, \Sigma)$. La simulation s'arrête à $t = 7$ et utilise un schéma d'Euler explicite avec un pas de temps de 0.01.

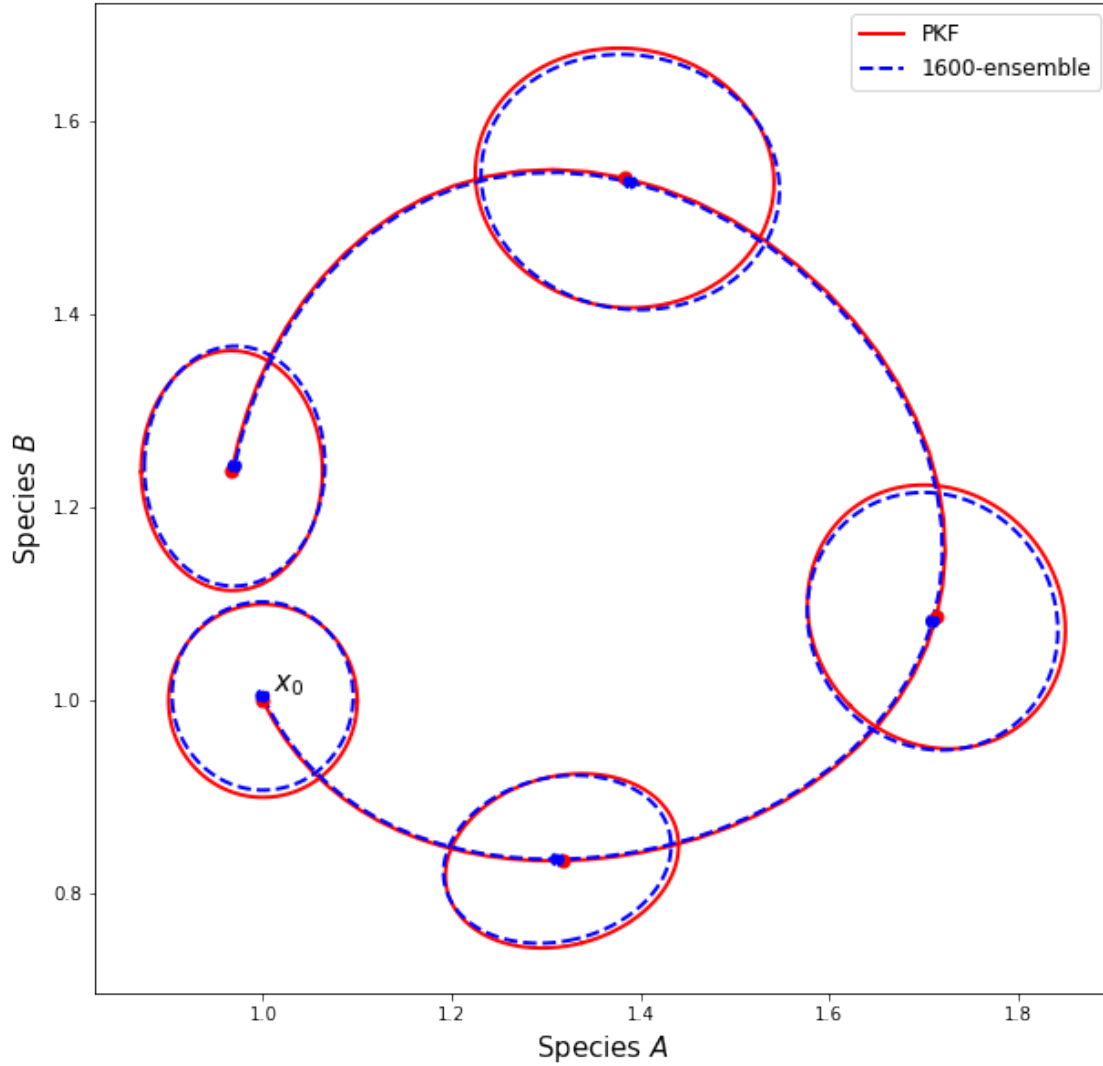


FIGURE 3 – PKF et EnKF - prévisions pour le système de Lotka-Volterra

Malgré les approximations faites dans la dynamique du système (8), le PKF est en mesure de fournir une prévision de l'état moyen et des statistiques d'erreurs parfaitement recevables. Notons qu'en plus d'être de qualité comparable à celle d'un grand ensemble, le PKF est dans ce cas environ 1600 fois moins coûteux en calculs.

3 Transport 1D d'un polluant dans l'atmosphère

A présent, nous nous intéressons au transport par le vent (noté u) d'un polluant A (une espèce chimique : *univarié*) dans l'atmosphère. On considère ici le cas d'une espèce chimique n'interagissant pas avec les autres composants chimique de l'atmosphère, ce peut être le cas par exemple des cendres volcaniques. On modélise ici l'espace dans lequel évolue ce polluant par le segment $]0, 1[$, que l'on périodise. La concentration du polluant A suit alors la dynamique suivante :

$$(Adv) : \partial_t A + u \partial_x A = -A \partial_x u \quad (9)$$

Disposant d'une connaissance *a priori* de l'état du système \mathbf{x}^b et de l'erreur lui étant associée Σ^b , il s'agit de fournir une prévision de l'état du système et l'erreur : \mathbf{x}^f et Σ^f . Considérant maintenant un problème spatialisé (1D) un nouveau paramètre est nécessaire au PKF afin de modéliser correctement les fonctions de covariance entre les points de l'espace. Ceci est fait en détail dans la partie suivante.

3.1 Formuler le PKF

Dans ce cadre 1D univarié, la modélisation des fonctions de covariance repose sur un champ de variance V_x , et un champ de longueur de portées L_x , tous deux définis en tout point \mathbf{x} du domaine. Le champ de variance V_x donne l'amplitude des erreurs en chaque point d'espace, et la longueur de portée elle caractérise l'étendue de la fonction de corrélation en un point de l'espace. On peut faire l'analogie entre la longueur de portée et le rôle de l'écart-type dans la forme d'une gaussienne. Notons qu'en dimension supérieure (2D, 3D), cette longueur de portée devient ce que l'on nomme le tenseur d'aspect noté s_x , lui même relié au tenseur métrique noté g_x par $g_x = s_x^{-1}$; là encore l'analogie avec la matrice de covariance d'une gaussienne multi-dimensionnelle peut-être faite pour appréhender le rôle du tenseur d'aspect. Le champs des tenseurs d'aspect est plus couramment désigné comme la "carte d'anisotropie". En 0D, la longueur de portée L_x est reliée au tenseur d'aspect (qui est en fait un scalaire) par $L_x = \sqrt{s_x}$. Nous verrons section ?? comment modéliser les fonctions de covariance à l'aide du champs de variance et de la carte d'anisotropie.

Le tenseur métrique, en dimension supérieure ou égale à 1, est relié au champ d'erreur ε [1] par :

$$g_{ij}(x) = \mathbb{E} \left[\partial_{x^i} \left(\frac{\varepsilon}{\sqrt{V}} \right) \partial_{x^j} \left(\frac{\varepsilon}{\sqrt{V}} \right) \right] (x) \quad (10)$$

On note parfois $\tilde{\varepsilon} = \varepsilon / \sqrt{V}$ l'erreur normalisée et $\sigma = \sqrt{V}$ l'écart-type. Dans le cas 1D, cette définition se réécrit :

$$g_x = \mathbb{E} \left[\left(\partial_x \left(\frac{\varepsilon_x}{\sqrt{V_x}} \right) \right)^2 \right]. \quad (11)$$

Pour ce problème de transport univarié, le PKF est alors constitué des paramètres d'état moyen, de champ de variance et de carte d'anisotropie : (A, V, s) . Afin de déterminer la dynamique de ces paramètres, nous utilisons le package SymPKF évoqué section 2.4.2.

$$(PKF\text{-Advection}) : \begin{cases} \partial_t A + u \partial_x A = -A \partial_x u & (12a) \\ \partial_t V + u \partial_x V = -2V \partial_x u & (12b) \\ \partial_t s + u \partial_x s = 2s \partial_x u & (12c) \end{cases}$$

Le système du PKF est entièrement découplé, et exact : aucun problème de fermeture. Il nous permettra de ce fait de mettre en perspective les biais qui peuvent apparaître dans les diagnostics

fournis par des ensembles lors des simulations numériques. Nous considérerons des ensemble de taille 1600, avec un bruit d'échantillonnage de l'ordre de 2.5% afin de comparer les deux méthodes lors d'une étape de prévision à horizon lointain.

3.2 Modèle de covariance univarié

Afin d'être en mesure d'assimiler des observations après une étape de prévision, il faut être en mesure d'estimer l'erreur associée à la prévision que l'on vient de réaliser. Dans le formalisme de l'EnKF, la matrice de covariance d'erreur de prévision est estimée empiriquement à partir de l'ensemble des prévisions $(x_i^f)_{i=1,\dots,N_e}$ où N_e est la taille de l'ensemble. On obtient alors une matrice Σ^f de taille $N_x \times N_x$ dans laquelle l'entrée $\Sigma_{i,j}^f$ correspond à la covariance prévue entre les points de grille x_i et x_j .

Dans le formalisme du PKF, cette covariance est estimée à l'aide d'un modèle de covariance qui repose sur le champs de variance V et la carte d'anisotropie s . D'après le théorème de Paciorek et Schervish [3], pour une fonction de corrélation homogène $\rho^h(\|\mathbf{x} - \mathbf{y}\|)$ dans \mathbb{R}^d de tenseur d'aspect \mathbf{I}_d et un champs différentiable de tenseurs définis postifs \mathbf{s} , il est possible de former une fonction de corrélation hétérogène définie par :

$$\rho(\mathbf{s})(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{s}_x|^{1/4} |\mathbf{s}_y|^{1/4}}{|\frac{1}{2}(\mathbf{s}_x + \mathbf{s}_y)|^{1/2}} \rho^h\left(\|\mathbf{x} - \mathbf{y}\|_{[\frac{1}{2}(\mathbf{s}_x + \mathbf{s}_y)]^{-1}}\right) \quad (13)$$

pour laquelle le tenseur d'aspect en \mathbf{x} est approximativement \mathbf{s}_x , et où $|\cdot|$ est le déterminant matriciel. Nous considérons une un modèle de corrélation gaussien, le modèle de covariance univarié s'écrit donc :

$$\mathbf{P}^{\text{he.gauss}}(V, \mathbf{s})(\mathbf{x}, \mathbf{y}) = \sqrt{V_x V_y} \frac{|\mathbf{s}_x|^{1/4} |\mathbf{s}_y|^{1/4}}{|\frac{1}{2}(\mathbf{s}_x + \mathbf{s}_y)|^{1/2}} \exp\left(-\frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_{[\frac{1}{2}(\mathbf{s}_x + \mathbf{s}_y)]^{-1}}^2\right) \quad (14)$$

A l'aide de ce modèle de covariance, nous sommes en mesure de fournir une approximation de la matrice de covariance associée à une prévision à l'aide des paramètres du PKF.

3.3 Simulation numérique

3.3.1 Paramètres de simulation

Paramètres communs à l'EnKF et au PKF : Initialement, on fixe de manière homogène le champs d'état moyen (ou concentration moyenne), de variance, et de longueur de portée à respectivement $A_0 = 1$, $V_0 = 0.125^2$ et $s_0 = l_h^2 = 0.08^2$. On considère que le segment $]0, 1[$ est une dilatation d'un segment $]0, D[$ où $D = 1000\text{km}$. Le champs de vent, constant dans le temps, est défini par $u(x) = (35 + 15 \cos(2\pi x))/D$, modélisant un vent d'intensité moyenne $35\text{km}\cdot\text{h}^{-1}$ avec une intensité maximale $U_{max} = 50\text{km}\cdot\text{h}^{-1}$ (Figure 4). On définit ensuite le temps caractéristique T comme étant le temps moyen nécessaire à un polluant pour effectuer une révolution autour du domaine, *i.e.* $T = D/\bar{u}$. L'horizon de prévision t_{end} est fixé à $3T$ ($\simeq 85\text{h}$).

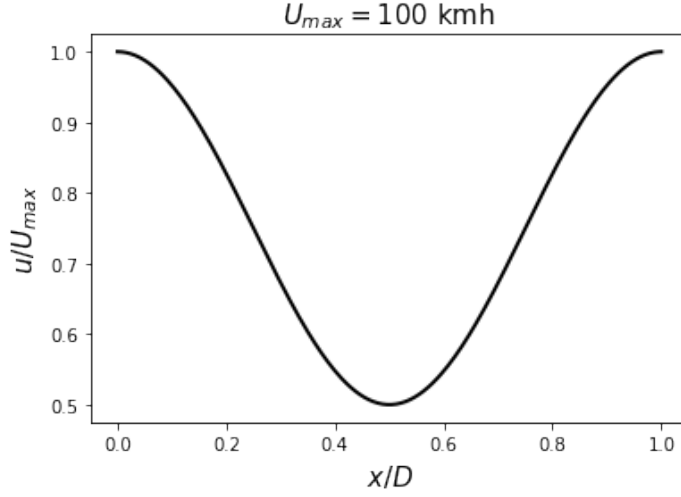


FIGURE 4 – Champs de vent

Schéma numérique employés : Le segment $]0, 1[$ est discrétisé en $N_x = 241$ points de grille. Les dérivés spatiales sont approchées par différences finies avec un schéma centré. Un schéma explicite Runge-Kutta d'ordre 4 (RK4) est utilisé pour l'intégration temporelle. Le pas de temps Δt considéré est déterminé par $\Delta t = \Delta x / U_{max}$, où $\Delta x = 1/N_x$ est le pas d'espace.

3.3.2 Génération des champs d'erreurs aléatoires

Afin de générer des champs aléatoires pour initialiser un ensemble, on procède comme suit. Etant donné les paramètres d'incertitude initiale V_0 et l_h , on définit la fonction de covariance ρ_h homogène par :

$$\rho_h(x, y) = V_0 \exp \left(-\frac{\|x - y\|_2^2}{2l_h^2} \right)$$

où l'on a périodisé la distance euclidienne $\|\cdot\|_2$ sur le segment $]0, 1[$. Soit ensuite N_x le nombre de points de grille de la discrétisation du segment $]0, 1[$, on définit la matrice \mathbf{P}^h par :

$$\mathbf{P}_{i,j}^h = \rho_h(x_i, x_j)$$

Enfin, on définit le champ aléatoire ε_h par :

$$\varepsilon_h = \mathbf{P}_h^{1/2} \varepsilon$$

où $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_x})$.

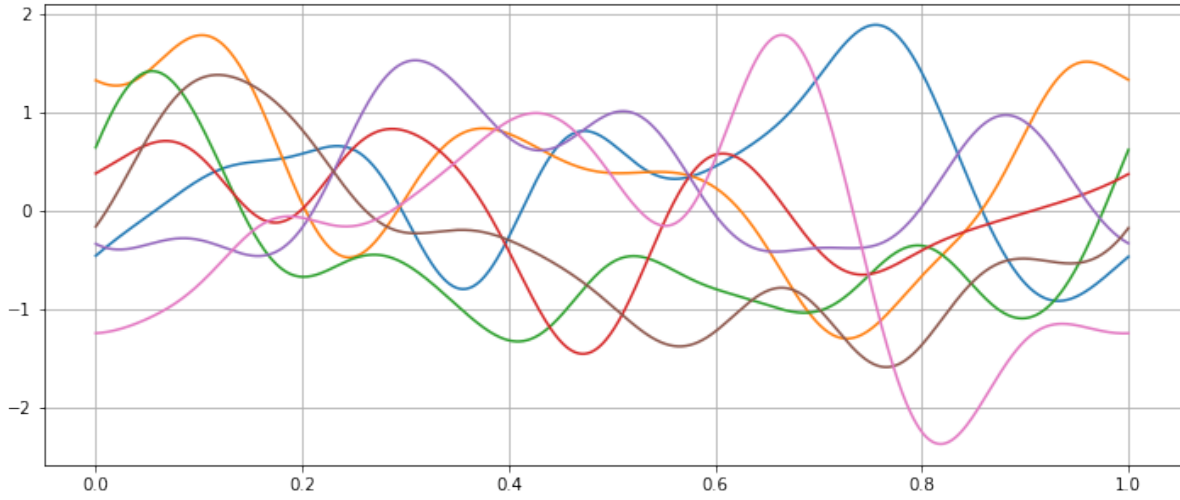


FIGURE 5 – Champs d’erreurs aléatoires

3.3.3 Résultats : comparaison des prévisions du PKF et de l’EnKF

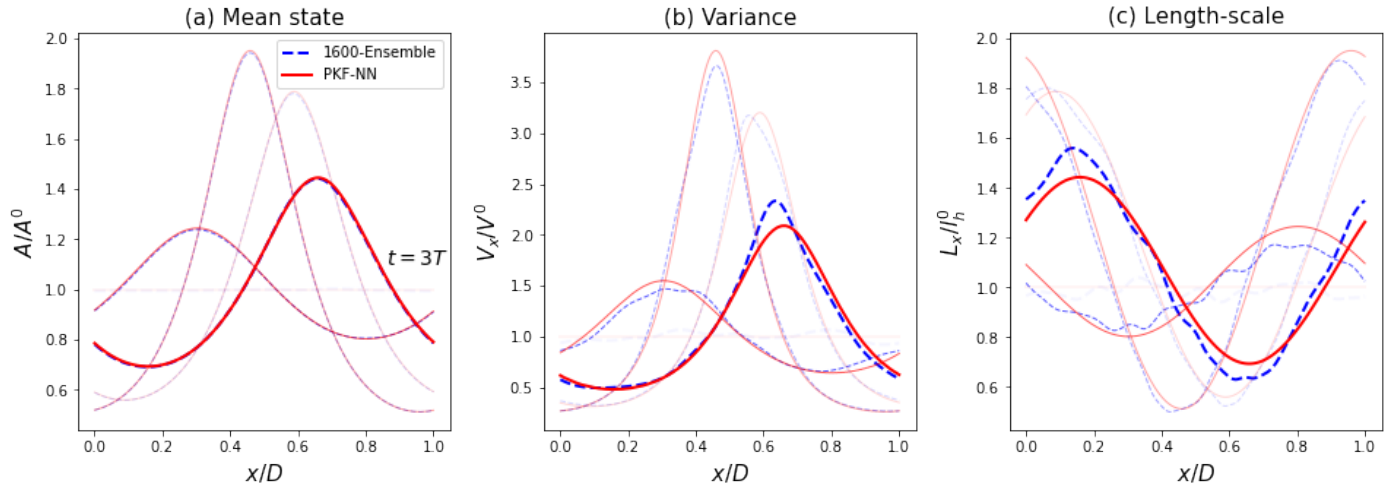


FIGURE 6 – Comparaison des prévisions du PKF et diagnostics de l’EnKF

Les deux méthodes fournissent des prévisions identiques similaires pour la concentration moyenne. Le champs de vent u que l’on s’est fixé plus haut entraîne une accumulation de la matière dans la région où le vent est le moins intense, le parallèle peut-être fait avec une situation d’embouteillage.

Les courbes de variance sont globalement identiques. L’incertitude croît dans le temps à tendance à se concentrer dans les zones d’accumulations de matière.

Enfin, pour les longueurs de portées, on remarque une plus grande sensibilité à l’erreur d’échantillonnage pour l’ensemble. Les résultats sont néanmoins les mêmes : les incertitudes sont plus localisées là où le vent est le moins intense, et inversement.

3.3.4 Résultats : comparaison des erreurs de prévisions pour le PKF et l'EnKF

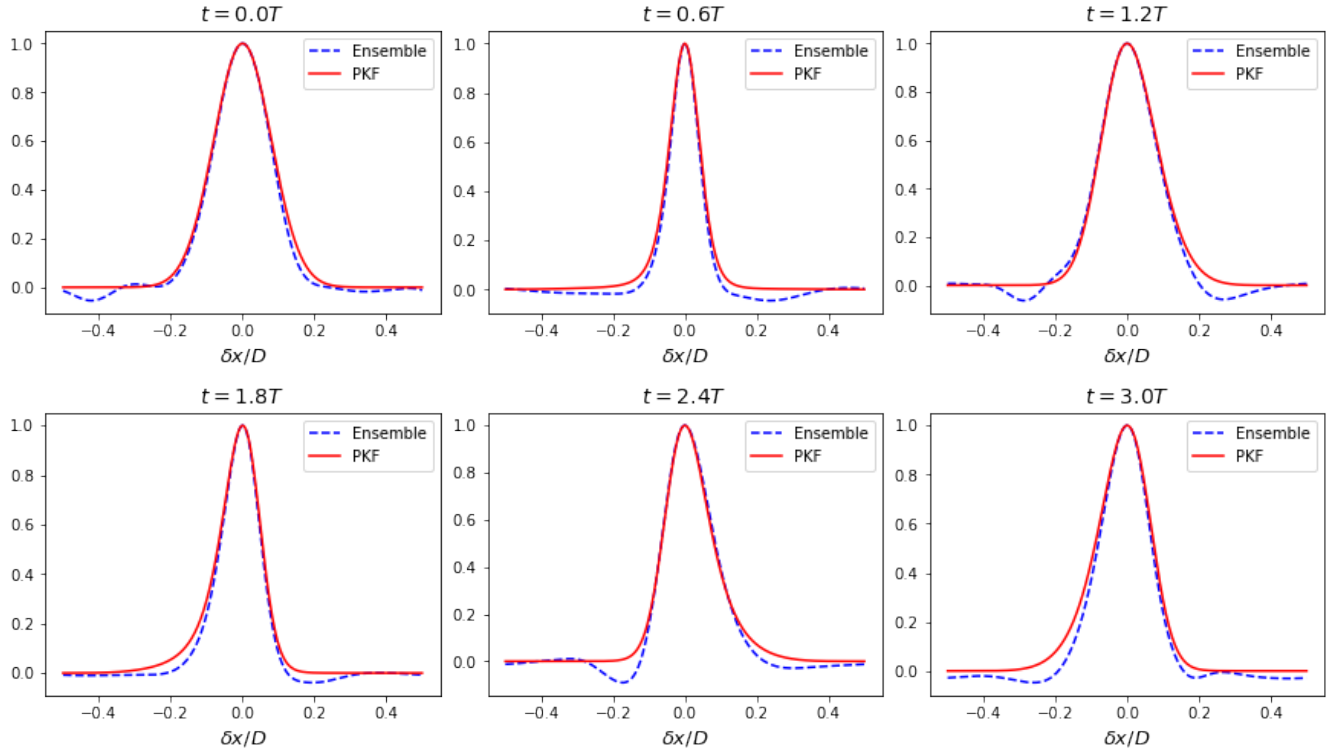


FIGURE 7 – Fonctions d'auto-corrélation à $x = 0.5$ à $t = [0.0, 0.6, 1.2, 1.8, 2.4, 3.0]T$

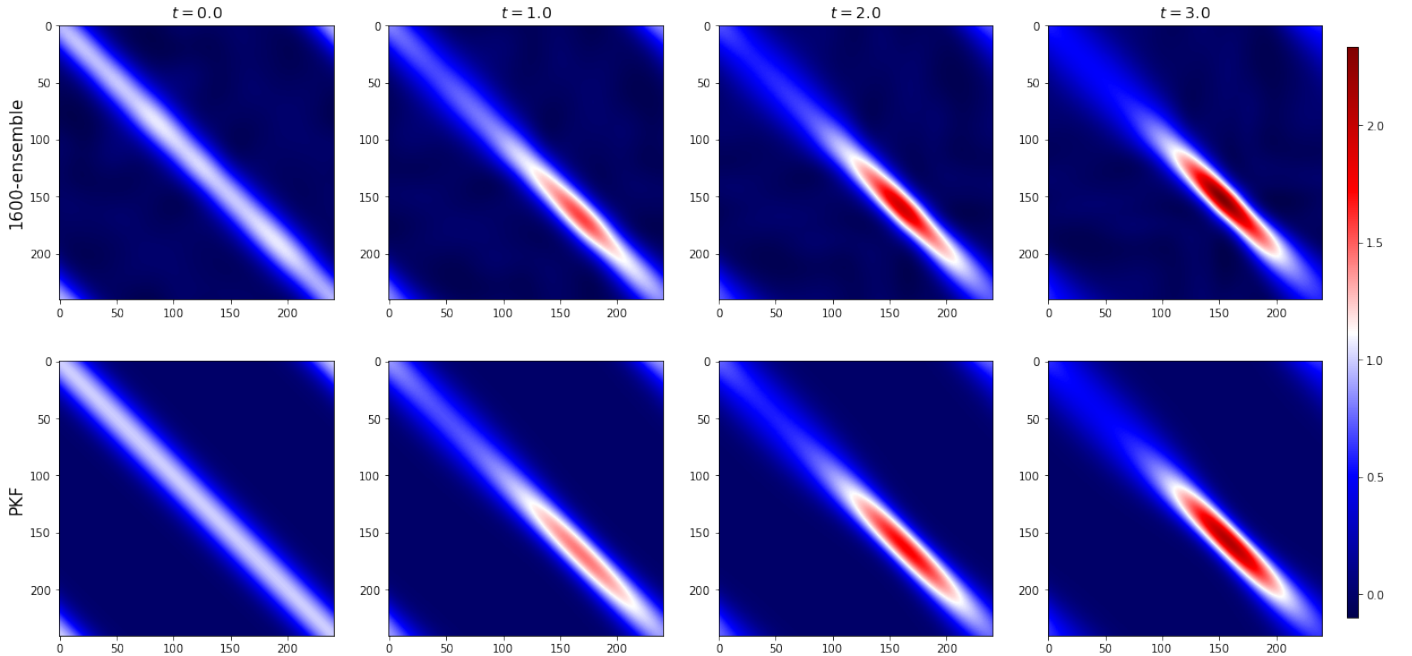
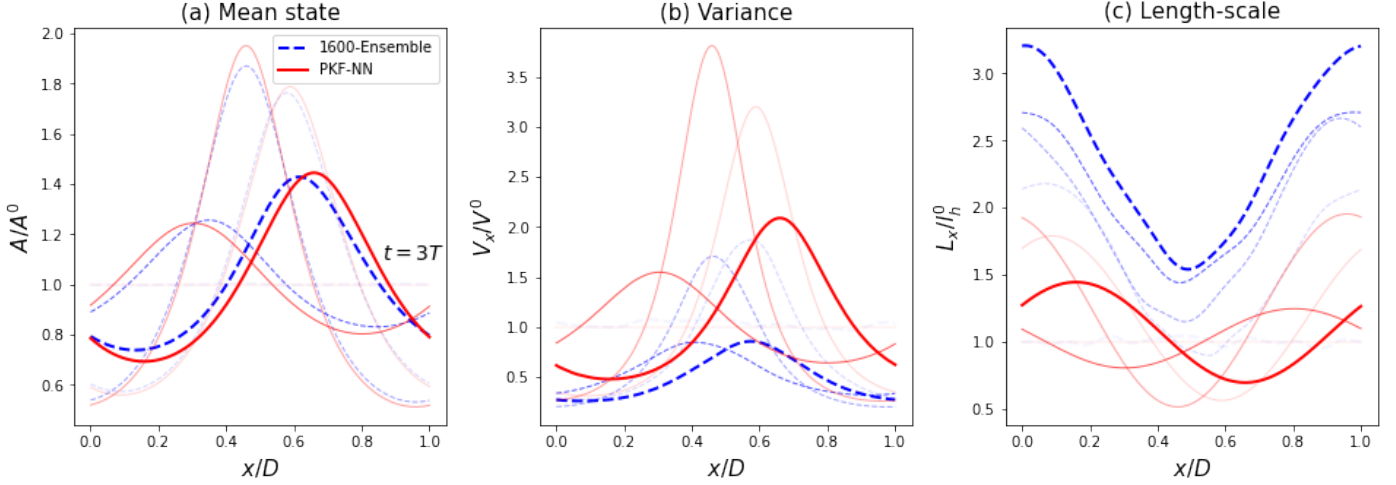


FIGURE 8 – Matrices de covariance d'erreur de prévision à $t = [0.0, 1.0, 2.0, 3.0]T$

3.4 Influence de l'erreur de modèle sur les diagnostics de l'EnKF

3.4.1 Mise en évidence d'un cas de divergence du filtre de Kalman d'ensemble

On souhaite ici mettre en avant l'influence que peut avoir l'erreur de modèle sur la qualité des diagnostics fournis par un ensemble. Les résultats obtenus précédemment contenaient relativement peu d'erreur : on avait quasi-superposition des courbes d'ensemble avec celles du PKF (exactes). Cela était dû au schéma d'ordre élevé (RK4) utilisé pour la prévision d'ensemble. Nous allons maintenant approximer les dérivés spatiales non plus par un schéma centré, mais par un schéma upwind, implicitement diffusif. On met ensuite en perspective les diagnostics de l'ensemble avec ceux fournis par le PKF.



On voit ici que le diagnostic obtenu pour la variance indique une incertitude significativement sous-estimée. Cela s'explique par la diffusion que comporte le schéma upwind explicite, qui a tendance à lisser les trajectoires, les rendant de plus en plus homogènes et ce faisant diminuant la variance. Dans le cas extrême où cette variance deviendrait nulle, on en viendrait à ne pas tenir compte des nouvelles observations lors de l'étape d'assimilation. On parle alors de divergence du filtre de Kalman d'ensemble. En d'autres termes, on "perd le signal".

3.4.2 Impact de la réduction des longueurs de portées initiales sur les diagnostics

Jusqu'ici on a considéré des longueurs de portées relativement grandes, $l_h = 0.08 \simeq 19\Delta x \simeq 80$ km pour les paramètres de simulations choisis. En considérant des longueurs de portées initiales plus petites, de l'ordre de $4\Delta x \simeq 16$ km, on est en mesure de décrire des phénomènes plus locaux mais cela vient générer beaucoup d'erreur de modèle comme nous allons le voir. On considère un schéma centré pour approcher les dérivées spatiales, et un schéma RK4 pour l'intégration temporelle. On met les résultats en perspective avec les prévisions du PKF, exactes quelque soient la longueur de portée initiale.

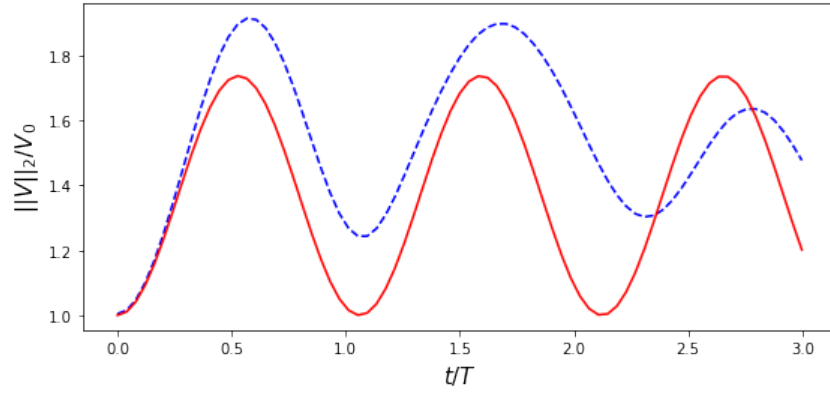


FIGURE 9 – Illustration du déphasage de la variance totale au cours du temps

3.5 Conclusion dans le cadre univarié de l'advection 1D

On peut se rendre compte au travers de cet exemple simple des avantages du PKF. D'abord, il permet à des coûts de calculs moindres (1600 fois inférieurs) l'obtention de diagnostics de meilleure qualité que l'ensemble car dépourvus d'erreurs d'échantillonnage. Ensuite, grâce à la modélisation des fonctions d'auto-corrélations, l'information normalement contenue dans une matrice de covariance gigantesque est résumée en quelques paramètres légers, nous dispensant du besoin d'une capacité immense de stockage. Enfin, on a vu que le PKF nous a permis de caractériser l'erreur de modèle associé à un schéma de discrétisation spatiale. De ce point de vue le PKF est un outil d'exploration offrant de nouveaux points de vue.

4 Un modèle de chimie transport simplifié pour l'exploration d'une formulation multivariée du PKF

A terme, on aimerait être capable de formuler un modèle de chimie-transport (CTM) reposant sur le formalisme du PKF. Un CTM consiste en la modélisation des interactions chimiques des espèces présentes dans l'atmosphère (120 espèces considérées dans MOCAGE) ainsi que de leur transport par le vent. Pour explorer la possibilité d'une formulation multivariée du PKF, on considère un modèle de chimie transport simplifié, qui contient néanmoins des composantes essentielles à un modèle plus réaliste (non-linéarités, plusieurs espèces, transport). Ce modèle consiste donc en la modélisation du transport dans l'atmosphère de deux espèces chimiques A et B dont les interactions chimiques sont modélisées par le système non-linéaire de Lotka-Volterra. Il s'agit donc de conjuguer les processus d'advection (9) et de Lotka-Volterra (6), donnant lieu au système suivant :

$$(\text{Adv-LV}) : \begin{cases} \partial_t A + u \partial_x A = -A \partial_x u + k_1 A - k_2 AB \\ \partial_t B + u \partial_x B = -B \partial_x u + k_2 AB - k_3 B \end{cases} \quad (15a)$$

$$(15b)$$

4.1 Formulation multivariée du PKF avec le SymPKF

Dans ce contexte, le système du PKF est constitué de sept paramètres : état moyen, champs de variance et carte d'anisotropie pour chaque espèce, ainsi qu'un champ de covariance propre au caractère multivarié. On détermine la dynamique des paramètres du PKF à l'aide du package SymPKF présenté dans la section (2.4.2). On illustre ici la simplicité d'obtention de la dynamique du PKF grâce au package.

FIGURE 10 – Définition d'un problème avec le SymPKF

```
from sympy import symbols, Function, Derivative
from sympkf.sympkf import PDESystem, Eq, SymbolicPKF, t

x = symbols("x")
k1, k2, k3 = symbols("k_1 k_2 k_3")
u = Function("u")(x) #wind
A = Function("A")(t, x) #tracer A
B = Function("B")(t, x) #tracer B
eA = Function("\varepsilon_{A}")(t, x) #tracer A
eB = Function("\varepsilon_{B}")(t, x) #tracer B
dynamics = [
    Eq(Derivative(A, t), -u*Derivative(A, x) - A*Derivative(u, x) + k1*A - k2*A*B),
    Eq(Derivative(B, t), -u*Derivative(B, x) - B*Derivative(u, x) - k3*B + k2*A*B)
]
display(dynamics[0]), display(dynamics[1]);3
```

$$\frac{\partial}{\partial t} A(t, x) = k_1 A(t, x) - k_2 A(t, x) B(t, x) - A(t, x) \frac{d}{dx} u(x) - u(x) \frac{\partial}{\partial x} A(t, x)$$

$$\frac{\partial}{\partial t} B(t, x) = k_2 A(t, x) B(t, x) - k_3 B(t, x) - B(t, x) \frac{d}{dx} u(x) - u(x) \frac{\partial}{\partial x} B(t, x)$$

Toute la partie fastidieuse et sujette à erreur de calcul à la main des dynamiques de chacune des quantités est alors traitée automatiquement par le SymPKF.

FIGURE 11 – Calcul de la dynamique des paramètres du PKF avec le SymPKF

```
system = PDESystem(dynamics)
pkf_system = SymbolicPKF(system)
for equation in pkf_system.in_aspect: display(equation)
```

$$\begin{aligned}
\frac{\partial}{\partial t} A(t, x) &= k_1 A(t, x) - k_2 A(t, x) B(t, x) - k_2 V_{AB}(t, x) - A(t, x) \frac{d}{dx} u(x) - u(x) \frac{\partial}{\partial x} A(t, x) \\
\frac{\partial}{\partial t} B(t, x) &= k_2 A(t, x) B(t, x) + k_2 V_{AB}(t, x) - k_3 B(t, x) - B(t, x) \frac{d}{dx} u(x) - u(x) \frac{\partial}{\partial x} B(t, x) \\
\frac{\partial}{\partial t} V_A(t, x) &= 2k_1 V_A(t, x) - 2k_2 A(t, x) V_{AB}(t, x) - 2k_2 B(t, x) V_A(t, x) - u(x) \frac{\partial}{\partial x} V_A(t, x) - 2 V_A(t, x) \frac{d}{dx} u(x) \\
\frac{\partial}{\partial t} V_B(t, x) &= 2k_2 A(t, x) V_B(t, x) + 2k_2 B(t, x) V_{AB}(t, x) - 2k_3 V_B(t, x) - u(x) \frac{\partial}{\partial x} V_B(t, x) - 2 V_B(t, x) \frac{d}{dx} u(x) \\
\frac{\partial}{\partial t} V_{AB}(t, x) &= k_1 V_{AB}(t, x) + k_2 A(t, x) V_{AB}(t, x) - k_2 A(t, x) V_B(t, x) - k_2 B(t, x) V_{AB}(t, x) + k_2 B(t, x) V_A(t, x) - k_3 V_{AB}(t, x) \\
&\quad - \frac{V_{AB}(t, x) u(x) \frac{\partial}{\partial x} V_B(t, x)}{2 V_B(t, x)} - \frac{V_{AB}(t, x) u(x) \frac{\partial}{\partial x} V_A(t, x)}{2 V_A(t, x)} - 2 V_{AB}(t, x) \frac{d}{dx} u(x) - u(x) \sqrt{V_A(t, x)} \sqrt{V_B(t, x)} \mathbb{E} \left(\varepsilon_A(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_B(t, x, \omega) \right) \\
&\quad - u(x) \sqrt{V_A(t, x)} \sqrt{V_B(t, x)} \mathbb{E} \left(\varepsilon_B(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_A(t, x, \omega) \right) \\
\frac{\partial}{\partial t} s_{A,xx}(t, x) &= - \frac{2k_2 A(t, x) V_{AB}(t, x) s_{A,xx}(t, x)}{V_A(t, x)} + \frac{2k_2 A(t, x) \sqrt{V_B(t, x)} s_{A,xx}^2(t, x) \mathbb{E} \left(\frac{\partial}{\partial x} \varepsilon_A(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_B(t, x, \omega) \right)}{\sqrt{V_A(t, x)}} \\
&\quad + \frac{k_2 A(t, x) s_{A,xx}^2(t, x) \mathbb{E} \left(\varepsilon_B(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_A(t, x, \omega) \right) \frac{\partial}{\partial x} V_B(t, x)}{\sqrt{V_A(t, x)} \sqrt{V_B(t, x)}} - \frac{k_2 A(t, x) \sqrt{V_B(t, x)} s_{A,xx}^2(t, x) \mathbb{E} \left(\varepsilon_B(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_A(t, x, \omega) \right) \frac{\partial}{\partial x} V_A(t, x)}{V_A^{\frac{3}{2}}(t, x)} \\
&\quad + \frac{2k_2 \sqrt{V_B(t, x)} s_{A,xx}^2(t, x) \mathbb{E} \left(\varepsilon_B(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_A(t, x, \omega) \right) \frac{\partial}{\partial x} A(t, x)}{\sqrt{V_A(t, x)}} - u(x) \frac{\partial}{\partial x} s_{A,xx}(t, x) + 2 s_{A,xx}(t, x) \frac{d}{dx} u(x) \\
\frac{\partial}{\partial t} s_{B,xx}(t, x) &= \frac{2k_2 B(t, x) V_{AB}(t, x) s_{B,xx}(t, x)}{V_B(t, x)} - \frac{2k_2 B(t, x) \sqrt{V_A(t, x)} s_{B,xx}^2(t, x) \mathbb{E} \left(\frac{\partial}{\partial x} \varepsilon_A(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_B(t, x, \omega) \right)}{\sqrt{V_B(t, x)}} \\
&\quad + \frac{k_2 B(t, x) \sqrt{V_A(t, x)} s_{B,xx}^2(t, x) \mathbb{E} \left(\varepsilon_A(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_B(t, x, \omega) \right) \frac{\partial}{\partial x} V_B(t, x)}{V_B^{\frac{3}{2}}(t, x)} - \frac{k_2 B(t, x) s_{B,xx}^2(t, x) \mathbb{E} \left(\varepsilon_A(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_B(t, x, \omega) \right) \frac{\partial}{\partial x} V_A(t, x)}{\sqrt{V_A(t, x)} \sqrt{V_B(t, x)}} \\
&\quad - \frac{2k_2 \sqrt{V_A(t, x)} s_{B,xx}^2(t, x) \mathbb{E} \left(\varepsilon_A(t, x, \omega) \frac{\partial}{\partial x} \varepsilon_B(t, x, \omega) \right) \frac{\partial}{\partial x} B(t, x)}{\sqrt{V_B(t, x)}} - u(x) \frac{\partial}{\partial x} s_{B,xx}(t, x) + 2 s_{B,xx}(t, x) \frac{d}{dx} u(x)
\end{aligned}$$

Le système obtenu se réécrit plus simplement :

$$\partial_t A + u\partial_x A = -A\partial_x u + k_1 A - k_2 AB - k_2 V_{AB} \quad (16a)$$

$$\partial_t B + u\partial_x B = -B\partial_x u - k_3 B + k_2 AB + k_2 V_{AB} \quad (16b)$$

$$\partial_t V_{AB} + u\partial_x V_{AB} = -2V_{AB}\partial_x u + V_{AB}(k_1 - k_2 B - k_3 + k_2 A) + k_2 V_A B - k_2 V_B A \quad (16c)$$

$$\partial_t V_A + u\partial_x V_A = -2V_A\partial_x u + 2[V_A(k_1 - k_2 B) - k_2 AV_{AB}] \quad (16d)$$

$$\partial_t V_B + u\partial_x V_B = -2V_B\partial_x u + 2[V_B(-k_3 + k_2 A) + k_2 BV_{AB}] \quad (16e)$$

$$\begin{aligned} \partial_t s_A = & \underbrace{-u\partial_x s_A + 2s_A\partial_x u}_{\text{advection}} - \frac{2k_2 s_A AV_{AB}}{V_A} + \frac{2k_2 As_A^2 \sigma_A}{\sigma_B} \overline{\partial_x \tilde{\varepsilon}_A \partial_x \tilde{\varepsilon}_B} + \\ & \overline{\tilde{\varepsilon}_B \partial_x \tilde{\varepsilon}_A} \left(-2k_2 As_A^2 \sigma_B \partial_x \sigma_A - \frac{2k_2 As_A^2 V_B \partial_x \sigma_B}{\sigma_A} + \frac{2k_2 \sigma_B s_A^2 \partial_x A}{\sigma_A} \right) \end{aligned} \quad (16f)$$

$$\begin{aligned} \partial_t s_B = & \underbrace{-u\partial_x s_B + 2s_B\partial_x u}_{\text{advection}} + \frac{2k_2 s_B BV_{AB}}{V_B} - \frac{2k_2 Bs_B^2 \sigma_A}{\sigma_B} \overline{\partial_x \tilde{\varepsilon}_A \partial_x \tilde{\varepsilon}_B} + \\ & \overline{\tilde{\varepsilon}_A \partial_x \tilde{\varepsilon}_B} \left(2k_2 Bs_B^2 \sigma_A \partial_x \sigma_B - \frac{2k_2 Bs_B^2 V_A \partial_x \sigma_A}{\sigma_B} - \frac{2k_2 \sigma_A s_B^2 \partial_x B}{\sigma_B} \right) \end{aligned} \quad (16g)$$

Plusieurs remarques. D'abord notons que pour les dynamiques (16a - 16e), il s'agit en fait du couplage du système PKF pour Lotka-Volterra (8) avec le PKF pour l'advection (12). La notation $\tilde{\varepsilon}$ correspond à l'erreur normalisée $\tilde{\varepsilon} = \varepsilon/\sigma$. Les termes en rouge sont exprimés en termes des erreurs et constituent donc des *problèmes de fermeture* pour la dynamique des tenseurs d'aspect (16f, 16g). Dans ces équations, on peut faire la distinction entre les termes liés uniquement au processus d'advection, comme dans le système (12), des autres termes, liés au modèle chimique. Nous voyons dans la section suivante comment simplifier la dynamique des tenseurs d'aspects afin de résoudre le problème de fermeture.

4.2 Simplification du système PKF

En diagnostiquant à partir d'une prévision d'ensemble les paramètres du PKF, l'on a observé (Figures ?? et 13) que les diagnostics des tenseurs d'aspects des deux espèces A et B étaient semblables à celui du tenseur d'aspect dans le cas univarié sans interaction chimique (Eq 12). Cette observation laisse à penser que les tenseurs d'aspects sont majoritairement mus par le processus d'advection.

Afin de vérifier cette intuition, on a donc à partir d'une prévision d'ensemble, diagnostiquer les trajectoires des tenseurs d'aspects, soit un ensemble de diagnostics $(s_A(t_k))_k$ et $(s_B(t_k))_k$ où $t_k = k * \Delta t$. Ces trajectoires nous permettent ensuite d'établir *a posteriori* la dynamique des tenseurs d'aspect $(\partial_t s_A(t_k))_k$ et $(\partial_t s_B(t_k))_k$ en procédant par différences finies. Enfin, on compare ces quantités aux deux composantes de la dynamique globale : celle liée à l'advection $(-u\partial_x s + 2s\partial_x u)$ et la partie restante liée à la chimie (différence entre globale et advection).

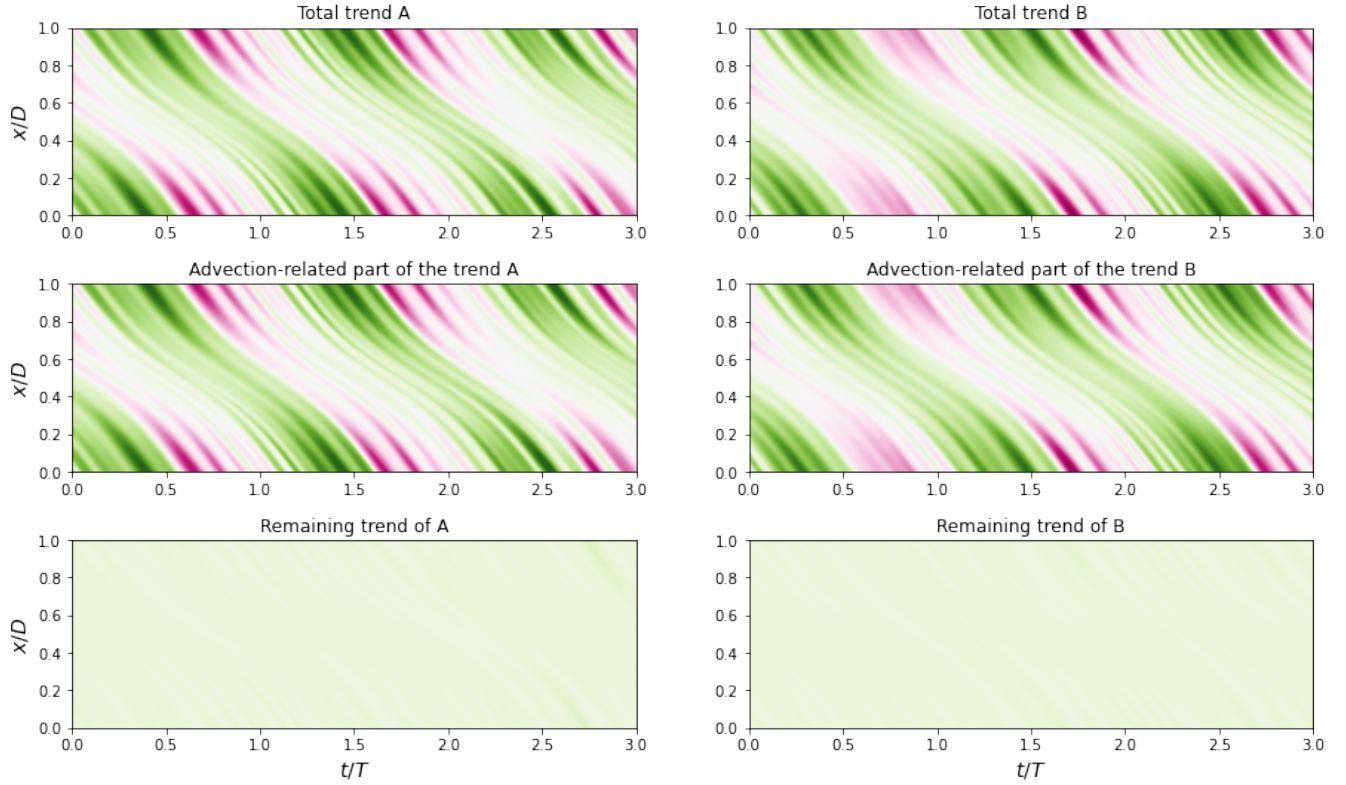


FIGURE 12 – Décomposition de la dynamique des tenseurs d’aspects - les échelles sont identiques pour chacun des graphiques.

On observe bien que l’advection explique en très grande partie la dynamique des tenseurs d’aspects. Ce résultat nous permet alors de simplifier grandement la dynamique des tenseurs d’aspect dans le système du PKF, pour donner lieu au système fermé suivant :

$$\text{PKF-Adv-LV : } \left\{ \begin{array}{ll} \partial_t A + u \partial_x A = -A \partial_x u + k_1 A - k_2 AB - k_2 V_{AB} & (17a) \\ \partial_t B + u \partial_x B = -B \partial_x u - k_3 B + k_2 AB + k_2 V_{AB} & (17b) \\ \partial_t V_{AB} + u \partial_x V_{AB} = -2V_{AB} \partial_x u + V_{AB}(k_1 - k_2 B - k_3 + k_2 A) + k_2 V_A B - k_2 V_B A & (17c) \\ \partial_t V_A + u \partial_x V_A = -2V_A \partial_x u + 2[V_A(k_1 - k_2 B) - k_2 A V_{AB}] & (17d) \\ \partial_t V_B + u \partial_x V_B = -2V_B \partial_x u + 2[V_B(-k_3 + k_2 A) + k_2 B V_{AB}] & (17e) \\ \partial_t s_A + u \partial_x s_A = 2s_A \partial_x u & (17f) \\ \partial_t s_B + u \partial_x s_B = 2s_B \partial_x u & (17g) \end{array} \right.$$

4.3 Simulation numérique : une prévision multivariée du système PKF

Le système du PKF (??) étant formulé pour (15), on peut désormais comparer les prévisions fournies par un ensemble avec celles du PKF.

4.3.1 Paramètres

paramètres du problème :

4.3.2 Paramètres de simulation

Les paramètres de simulations sont identiques à ceux du problème d'advection (section ??) à l'exception de ce qui relève du caractère bivarié. Les erreurs initiales ε_0 se décomposent maintenant comme deux champs d'erreurs $\varepsilon_0 = (\varepsilon_0^A, \varepsilon_0^B)$. On considère que ces deux erreurs sont décorrélées, *ie* $V_0^{AB} = 0$, et qu'elles sont initialisées de la même manière que dans le problème d'advection.

4.3.3 Résultats : comparaison des prévisions PKF et EnKF

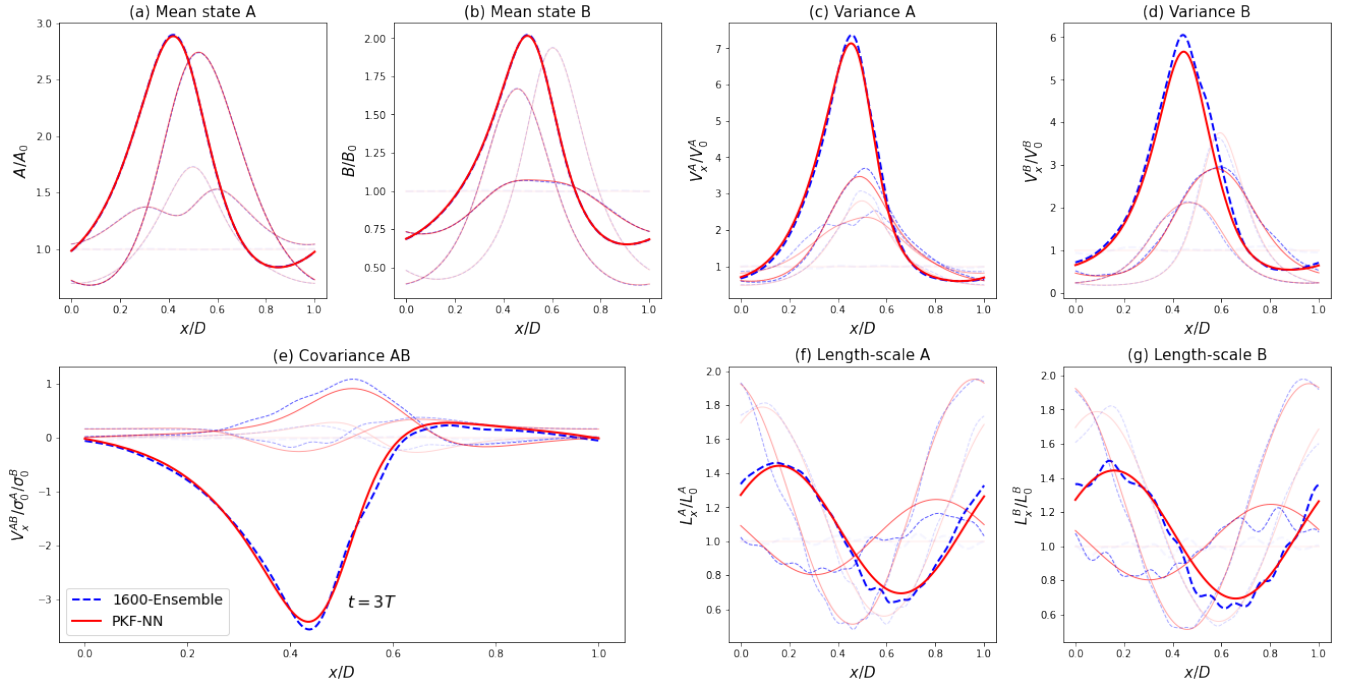


FIGURE 13 – Prévisions pour le système (15). PKF (lignes continues rouge) et EnKF (lignes pointillées bleu).

4.3.4 Résultats : comparaison des erreurs de prévisions PKF et EnKF

4.4 Définition d'un modèle de corrélation multivarié

Dans cette sous-section, on s'appuiera sur les diagnostics fournis par un ensemble afin de déterminer un modèle de corrélations croisées.

4.4.1 Diagnostics à l'aide d'un ensemble

On simule un ensemble de taille 1600 (2.5% d'erreur), dont les champs initiaux A et B ont pour moyenne 1, avec un bruit initial d'écart-type $\sigma_0 = 0.125$ et pour longueur de portée $lh_0 = 0.08$. On arrête la simulation à $t = 5$. On affiche figure 13 les diagnostics normalisés des différents paramètres.

On fait également les diagnostics des matrices de corrélations croisées à différents temps, visibles figure 14. Mis à part le bruit d'échantillonnage, à $t = 0$ les incertitudes sur les espèces sont décorrélées. Ensuite, la dynamique du système prend le dessus et les corrélations apparaissent. D'abord on remarque la non stationnarité de ces matrices de corrélations due au mouvement par le vent u , ce qui est assez intuitif : le vent variant le long du segment modifie les longueurs de portées. Puis, le caractère non symétrique de ces matrices. On cherche donc à fournir une approximation de la fonction de corrélation inter-espèces définie par :

$$\rho_{AB}(x, y) = \frac{\overline{\varepsilon_x^A \varepsilon_y^B}}{\sigma_x^A \sigma_y^B}$$

FIGURE 14 – Diagnostics des paramètres du PKF via un ensemble

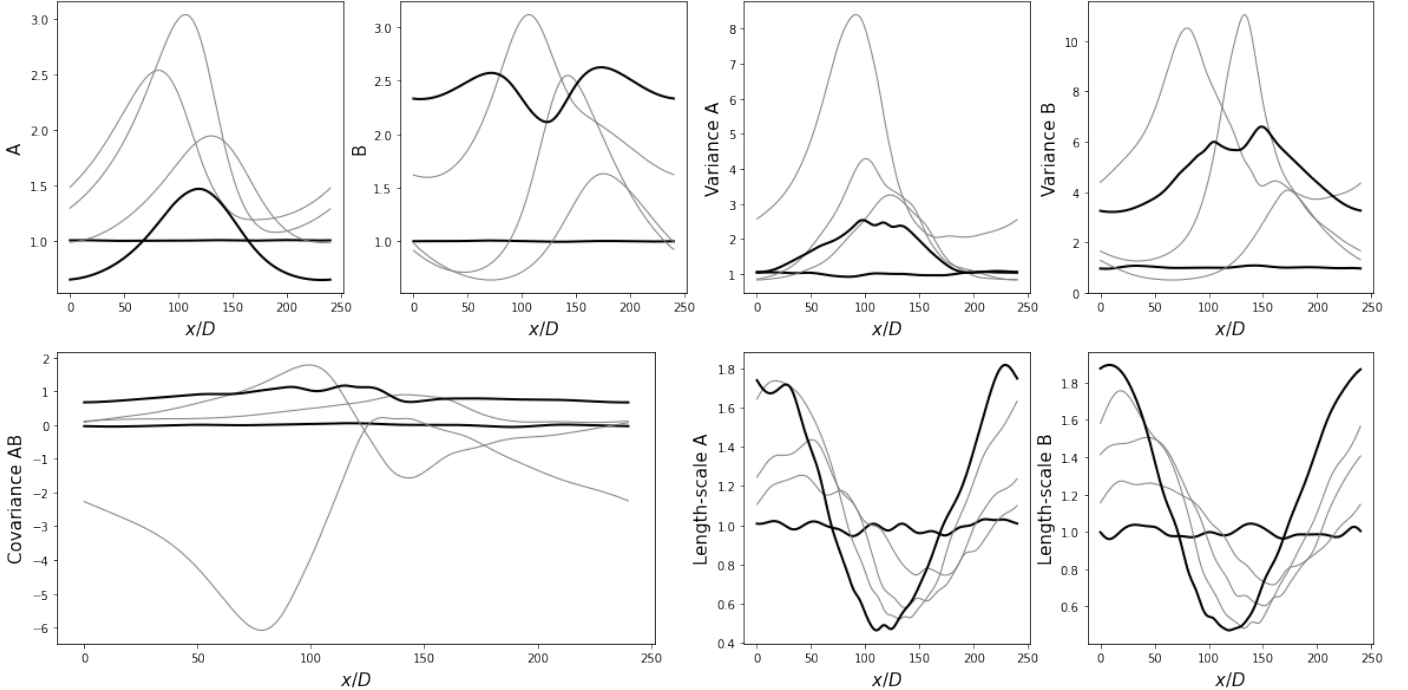
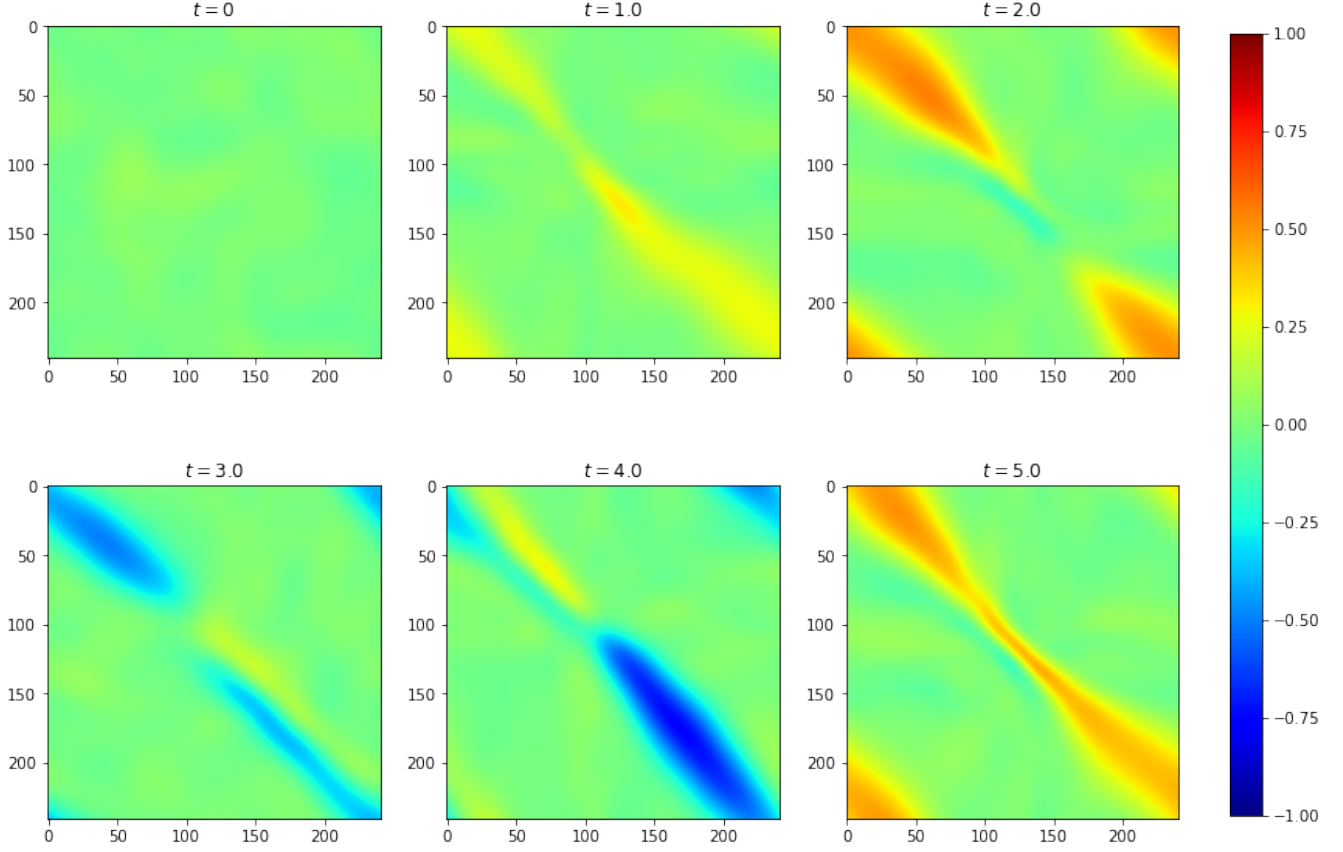


FIGURE 15 – Dynamique des corrélations croisées



4.4.2 Une première formulation analytique

La corrélation inter-espèce en un point x , $\rho_{AB}(x, x)$ est fournie par la relation :

$$\rho_{AB}(x, x) = \frac{V_x^{AB}}{\sigma_x^A \sigma_x^B}$$

Cela nous conduit à proposer la formulation suivante :

$$\hat{\rho}_{AB}(x, x + \delta x) = \frac{V_x^{AB}}{\sigma_x^A \sigma_x^B} \exp \left(-\frac{\delta x^2}{(s_x^A + s_x^B + s_{x+\delta x}^A + s_{x+\delta x}^B)/2} \right)$$

On montre en suivant quelques exemples de reconstructions des fonctions de corrélations ainsi que des matrices de corrélations croisées.

FIGURE 16 – Approximation analytique des corrélations croisées - matrices

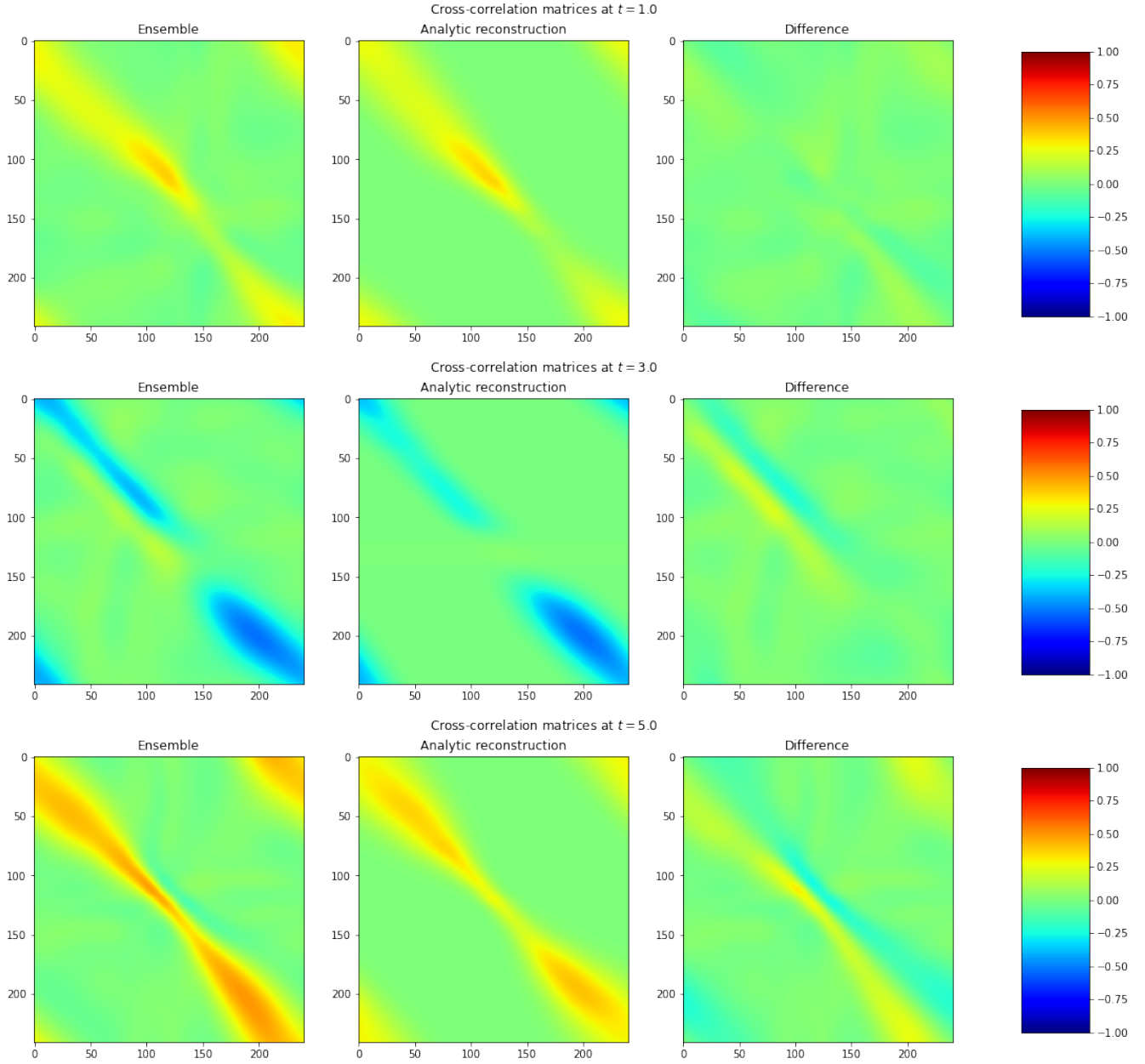
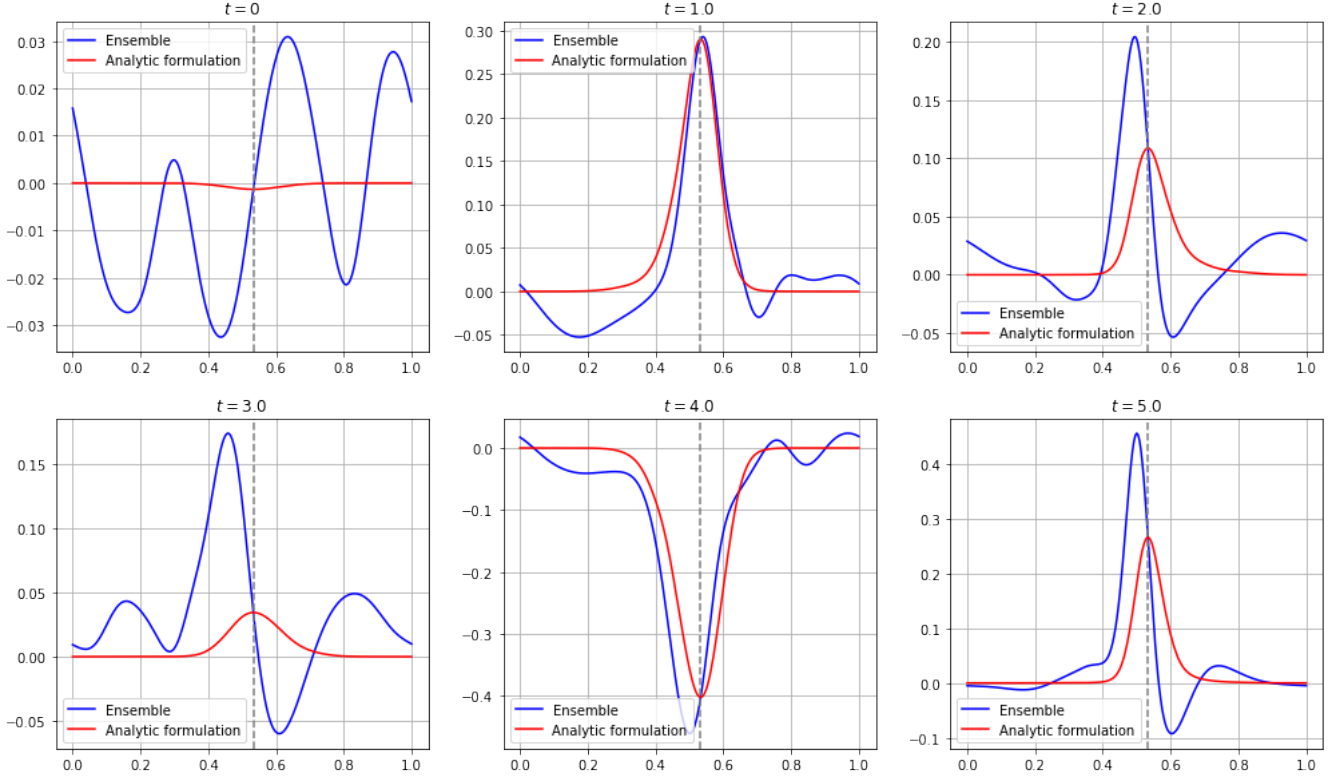


FIGURE 17 – Approximation analytique des corrélations croisées - courbes à $x = 0.5$



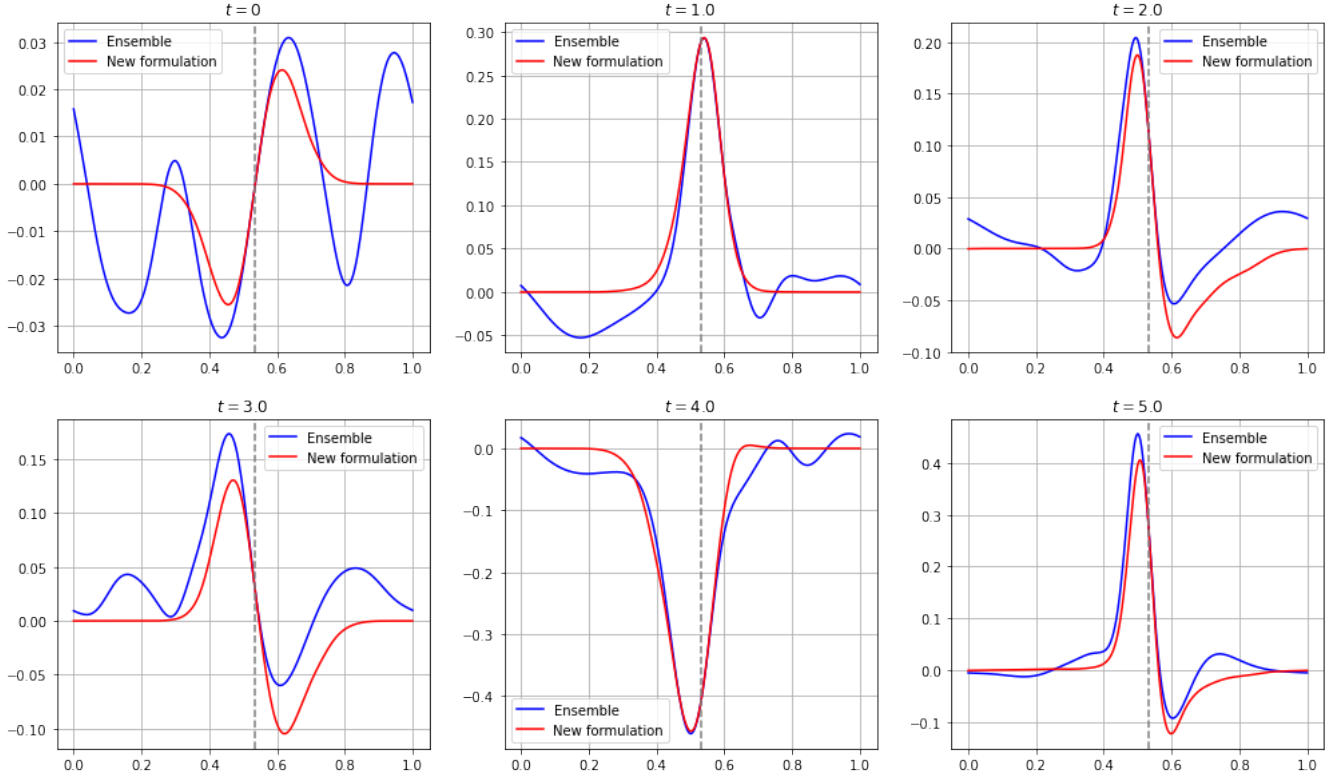
Les résultats fournis par cette approximations ne permettent pas de capturer l'aspect asymétrique des courbes à approximer. Cela est particulièrement visibles figure 15 à $t = 3$ et $t = 5$ ainsi que figure 16. Néanmoins, les amplitudes semblent correctes ainsi que les longueurs de portées. Il nous faut donc chercher à mieux représenter la composante asymétrique des fonctions de corrélations croisées.

4.4.3 Une voie d'amélioration du modèle de corrélations croisées

Si nous disposons de la quantité $\partial_y \rho_{AB}(x, x)$, alors nous serions en capacité de proposer la formulation suivante pour l'approximation des fonctions de corrélations croisées :

$$\rho_{AB}(x, x + \delta x) \simeq \left(\frac{V_x^{AB}}{\sigma_x^A \sigma_x^B} + \partial_y \rho_{AB}(x, x) \delta x \right) \exp \left(- \frac{\delta x^2}{(s_x^A + s_x^B + s_{x+\delta x}^A + s_{x+\delta x}^B)/2} \right)$$

FIGURE 18 – Performances nouvelle formulation - courbes



Cette formulation, ici exacte à l'ordre 1 car s'appuyant sur une quantité dont nous ne sommes normalement pas supposés disposer, se montre très encline à corriger le manque d'asymétrie de la formulation précédente.

4.4.4 Un modèle de corrélations croisées hybride : analytique et réseau de neurones

Comme vu précédemment, disposer d'une approximation de la quantité $\partial_y \rho_{AB}(x, x)$ permettrait l'accès à un meilleur modèle de corrélation croisées. Pour ce faire, on utilise un réseau de neurones.

Constitution d'une base de données : On a généré un grand nombre d'ensembles de taille 1600 et dont on a fait varier les champs initiaux A et B ainsi que les champs de vent. On a récupéré à différents temps les diagnostics des variables d'inférences $X : (A, B, V_{AB}, V_A, V_B, s_A, s_B, u)$ et ceux des quantités à estimer $Y : \partial_y \rho_{AB}(x, x)$.

Ajout de features : afin d'aider notre modèle à mieux estimer Y , on ajoute un certain nombre de features, comme par exemple les dérivées en espace des paramètres, ou encore la dérivée en espace du coefficient de corrélation. En effet, si l'on regarde une matrice de corrélation croisées, nous cherchons à estimer la dérivée selon y le long de la diagonale. Les paramètres du PKF nous permettent déjà de connaître la dérivée directionnelle selon $x = y$, c'est donc un paramètre qui semble important.

Architecture du réseau de neurones : En entrée on a un vecteur au format (26,241), 26 étant le nombre de features et 241 le nombre de points de grilles de notre discrétisation. On applique deux fois le cycle suivant : périodisation - convolution - maxpooling. En sortie on a pour format(1,241), on choisit comme fonction de perte l'erreur quadratique moyenne car il s'agit ici d'un problème de régression.

Un modèle hybride notre modèle de covariance est donc le suivant :

$$\hat{\rho}_{AB}(x, x + \delta x) = \left(\frac{V_x^{AB}}{\sigma_x^A \sigma_x^B} + NN(X) \right) \exp \left(- \frac{\delta x^2}{(s_x^A + s_x^B + s_{x+\delta x}^A + s_{x+\delta x}^B)/2} \right)$$

avec NN le réseau de neurones et X le vecteur contenant les paramètres du PKF + le vent.

Résultats : On montre en suivant les résultats que fournis ce modèle de corrélations croisées

FIGURE 19 – Modèle hybride pour les corrélations croisées - matrices

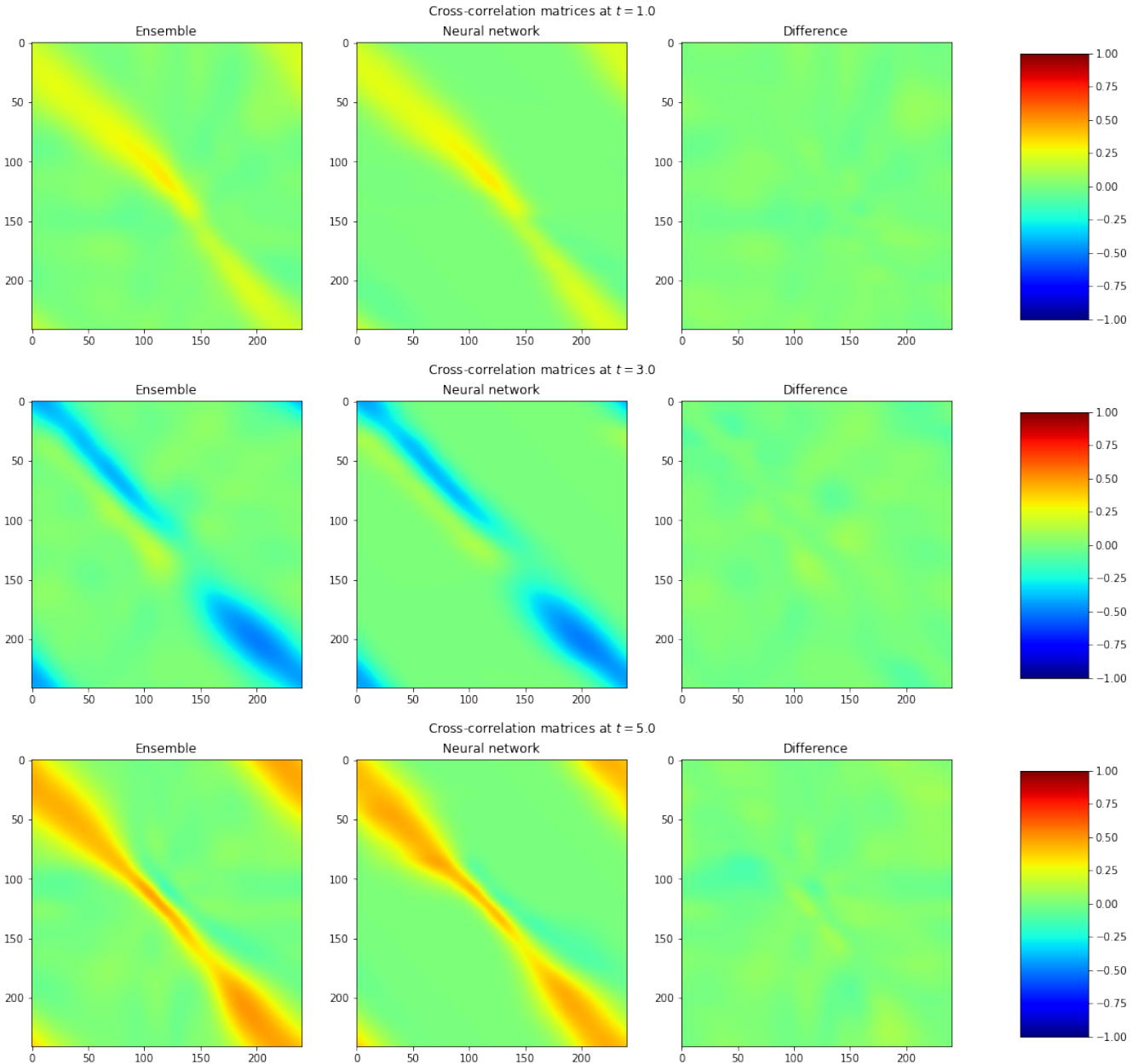
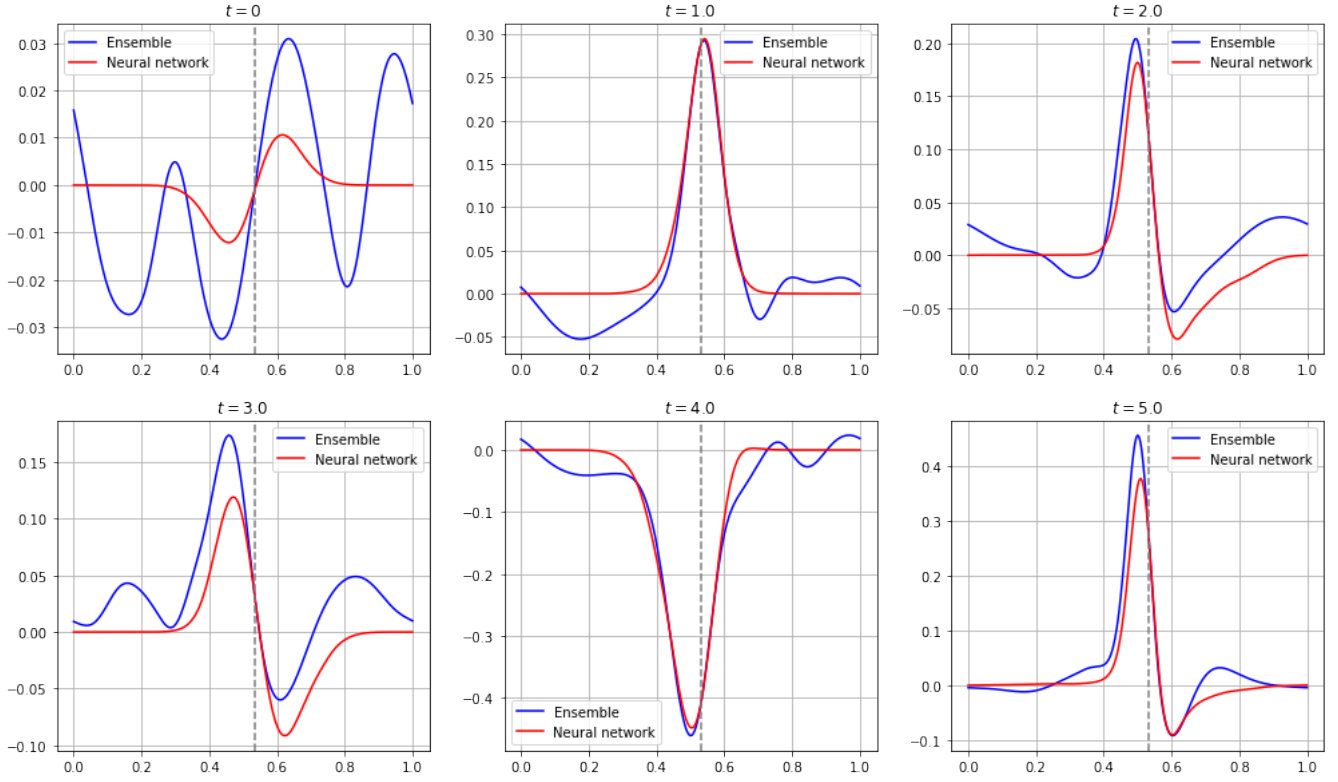


FIGURE 20 – Modèle hybride pour les corrélations croisées - courbes à $x = 0.5$



Conclusion : On constate que les pentes estimées par le réseau de neurones sont vraiment bonnes par rapport à ce que l'on doit observer. Cette formulation nous permet même de représenter des comportements plus complexes ($t = 2, 3, 5$). Les estimations semblent globalement correctes. Il faut aussi prendre en compte le fait que les courbes des ensembles sont des références, mais seulement dans une certaine mesure, car sujettes au bruit d'échantillonnage. Ainsi investiguer davantage dans la recherche d'un modèle de covariance qui collerait davantage à ces courbes ne fait peut-être pas sens.

4.5 Assimilation d'une observation : PKF et EnKF

5 Vers un modèle chimique transport plus réaliste

Afin de gagner en réalisme avec le PKF, on s'intéresse à un modèle chimique (dans un cadre 0D) plus réaliste et représentant des dynamiques plus complexes que celle qui peuvent être décrites par le système de Lotka-Volterra (6).

5.1 Introduction du modèle chimique GRS

Le modèle auquel on s'intéresse ici, intitulé GRS [?] [?] pour *generic reaction set*, décrit la dynamique d'un nombre réduit d'espèces ou méta-espèces chimiques présentes dans l'atmosphère.

Les réactions chimiques décrites par le modèle GRS sont les suivantes :



où ROC, RP et S(N)GN signifient respectivement *Reactive Organic Compound*, *Radical Pool* et *Stable (Non-) Gaseous Nitrogen product*.

Les émissions dans l'atmosphère de chacune de ces espèces sont également considérées, ainsi que leur dépôt au sol à un taux commun λ . Le système d'équations différentielles s'écrit ainsi :

$$(GRS) : \begin{cases} \frac{d[ROC]}{dt} = -\lambda[ROC] + E_{ROC} & (19a) \\ \frac{d[RP]}{dt} = -\lambda[RP] + k_1[ROC] - [RP] (k_2[NO] + 2k_6[NO_2] + k_5[RP]) & (19b) \\ \frac{d[NO]}{dt} = -\lambda[NO] + E_{NO} + k_3[NO_2] - [NO] (k_2[RP] + k_4[O_3]) & (19c) \\ \frac{d[NO_2]}{dt} = -\lambda[NO_2] + E_{NO_2} + k_4[NO][O_3] + k_2[NO][RP] - [NO_2] (k_3 + 2k_6[RP]) & (19d) \\ \frac{d[O_3]}{dt} = -\lambda[O_3] + k_3[NO_2] - k_4[NO][O_3] & (19e) \\ \frac{d[S(N)GN]}{dt} = -\lambda[S(N)GN] + 2k_6[NO_2][RP] & (19f) \end{cases}$$

La vitesse de réaction k_1 est reliée à la vitesse de photolyse k_3 par $k_1 = 0.152k_3$. Cette vitesse de photolyse varie au cours d'une journée (voir figure 22).

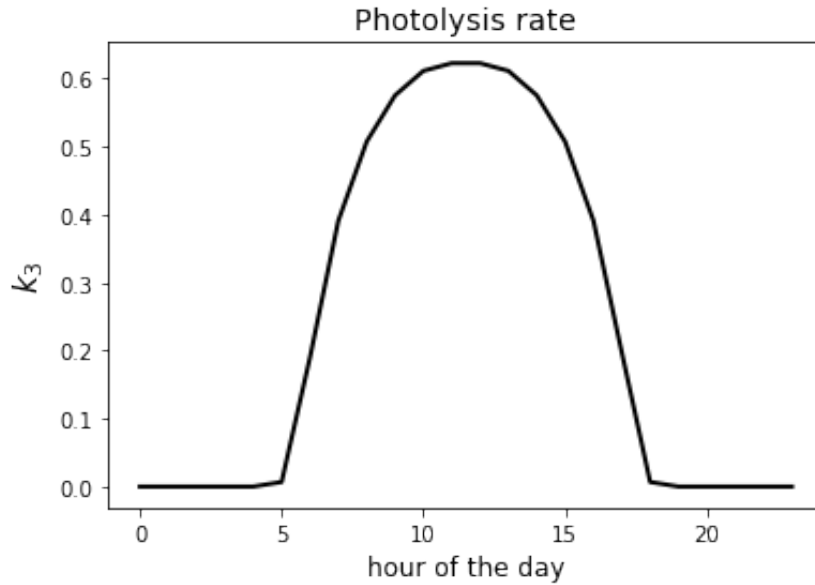


FIGURE 21 – Valeurs de k_3 aucours d'une journée

Toutes les valeurs des paramètres chimiques (concentrations initiales, facteurs d'émissions, vitesses de réaction ...) sont tirées de [?].

5.2 Comparaison prévision PKF et EnKF

On ne décrit pas ici l'étape de formulation du PKF (réalisée à l'aide du SymPKF). On rappelle qu'il est constitué dans ce contexte des états moyens et variances de chaque espèce ($6 + 6$ paramètres) ainsi des covariances de toutes les paires d'espèces (15 paramètres), soit 27 paramètres au total.

génération d'un état et d'une erreur initiale : Toutes les concentrations étant initialement fixées à zéro dans l'article, il n'est pas possible de générer un ensemble. On lance donc une prévision partant de zéro, stoppons la simulation dès que toutes les espèces sont en concentrations non nulles et décrétons une erreur de l'ordre de 10% des concentrations observées. On considère que les espèces sont indépendantes les unes des autres à ce stade. Ces paramètres constitueront nos données initiales pour le PKF et l'EnKF.

paramètres de simulation numérique : On considère un schéma d'Euler explicite avec un pas de temps $\Delta t = 0.001$. On arrête la simulation à $T = 72h$.

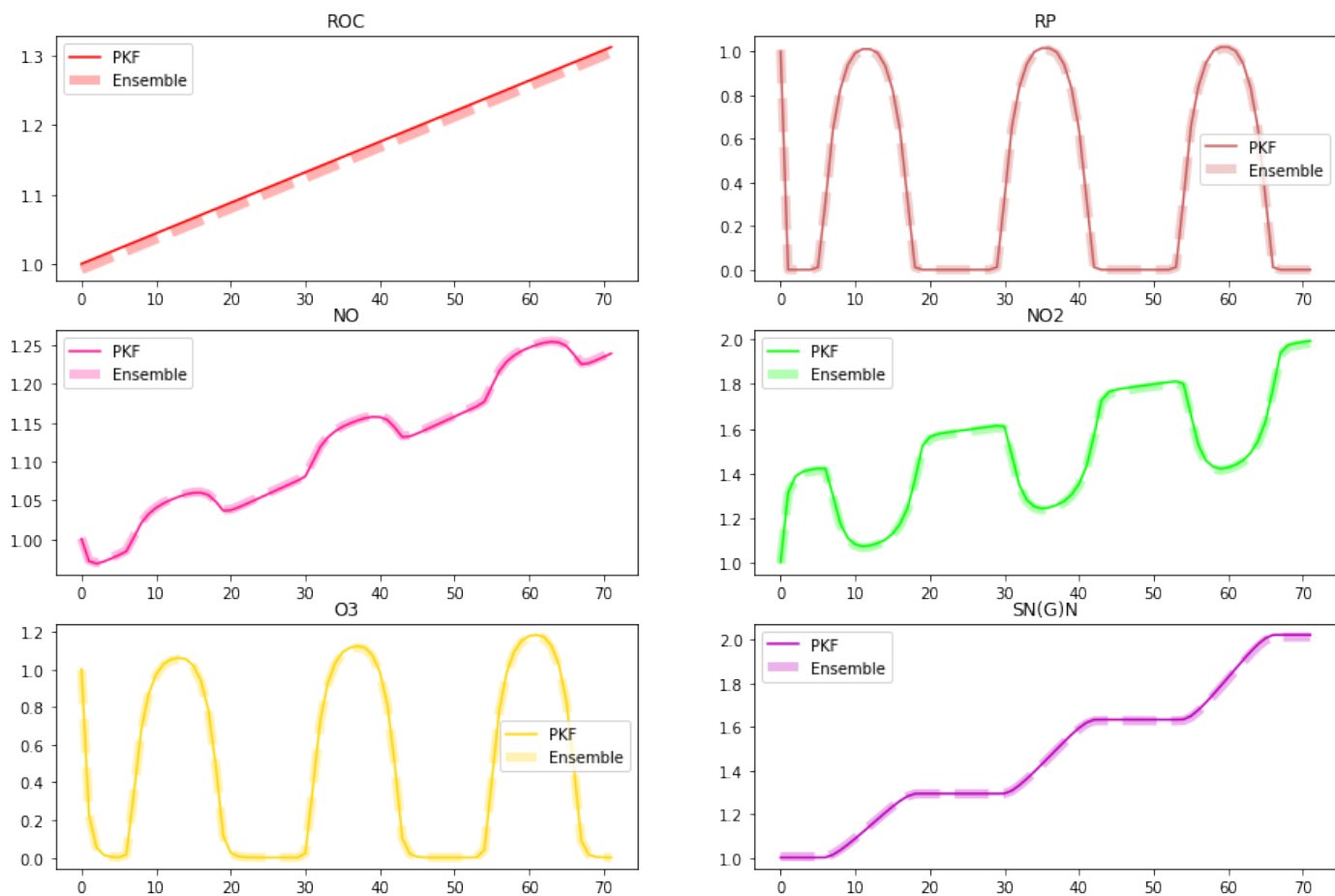


FIGURE 22 – Prédiction des concentrations. PKF (lignes continues et EnKF (lignes épaisses en pointillés). Les prévisions sont normalisées par les concentrations initiales.

On observe que pour les deux méthodes les prévisions sont identiques. Est visible également l'impact de la variation du taux de photolyse k_3 au long d'une journée.

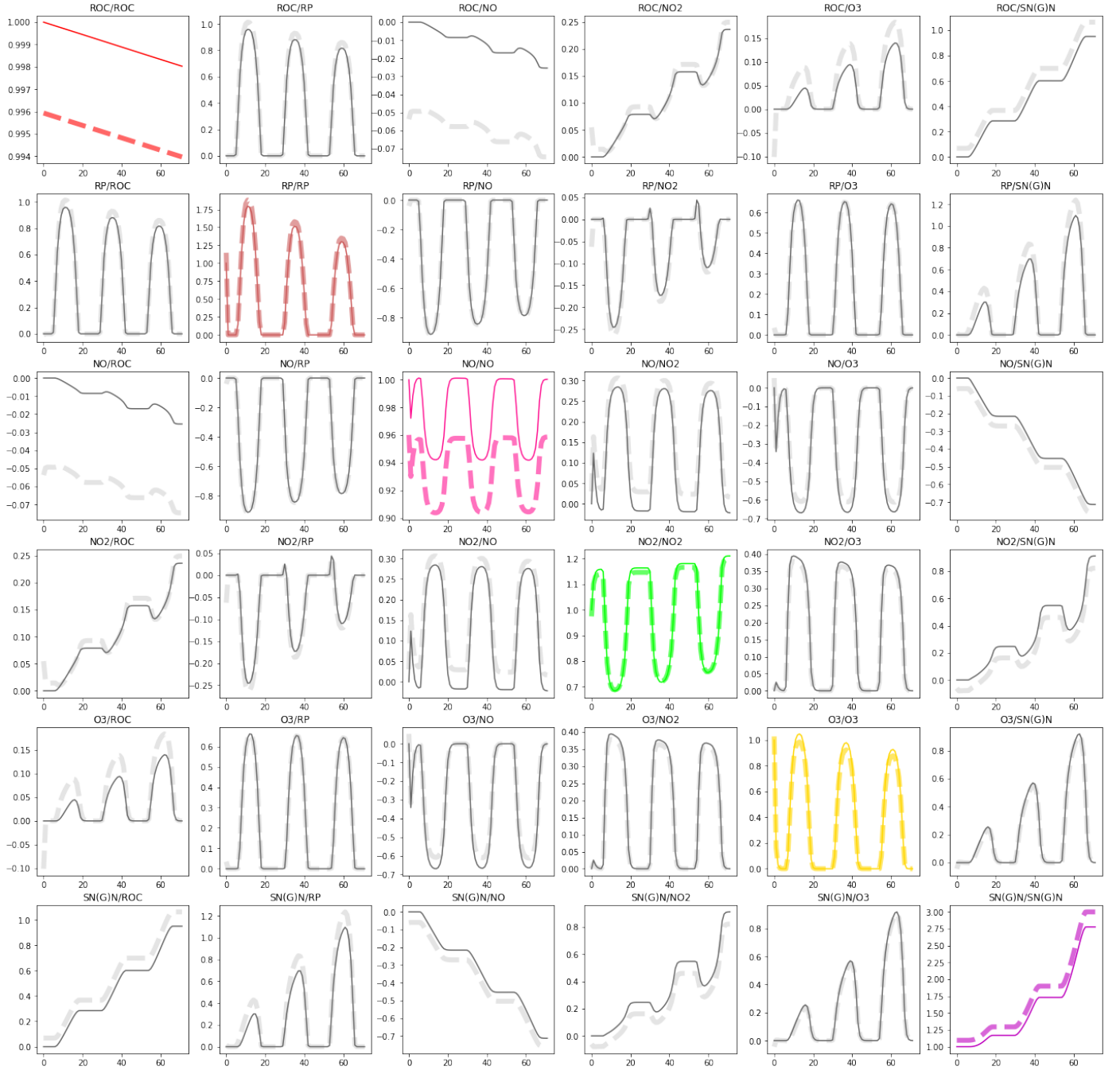


FIGURE 23 – Matrices de covariances d’erreurs de prévision. PKF (ligne continue) et EnKF (ligne épaisse en pointillés). Les variances sont normalisées leur valeur initiale, les covariances par les écart-types des espèces correspondantes.

On observe que le PKF décrit parfaitement la dynamique des incertitudes : les tendances sont les mêmes pour les deux méthodes, le bruit d’échantillonnage en moins pour le PKF.

Ce résultat est motivant : dans un cadre bien plus complexe que celui de Lotka-Volterra (6), le PKF parvient à décrire la dynamique des incertitudes. En sachant que le processus d’advection est parfaitement décrit par le PKF (3.1), il est raisonnable de penser que le PKF puisse décrire la dynamique des incertitudes dans un modèle combinant GRS et transport, comme on l’a vu dans la section 4.

Références

- [1] Daley. *Atmospheric Data Analysis*. 1991.
- [2] R. E. Kalman and Others. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1) :35–45, 1960.
- [3] C. Paciorek and M. Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in Neural Information Processing Systems*, 16, 03 2004.
- [4] O. Pannekoucke and P. Arbogast. Sympkf : a symbolic and computational toolbox for the design of parametric kalman filter dynamics. Mar. 2021.
- [5] O. Pannekoucke, S. Ricci, S. Barthelemy, R. Ménard, and O. Thual. Parametric kalman filter for chemical transport models. *Tellus A : Dynamic Meteorology and Oceanography*, 68(1) :31547, 2016. doi : 10.3402/tellusa.v68.31547. URL <https://doi.org/10.3402/tellusa.v68.31547>.