Prévision d'ensemble

Notes de cours (ENSTA, B10-2)

Vivien Mallet
INRIA*
Université Paris-Est†
vivien.mallet@inria.fr

Janvier 2008

Table des matières

*Équipe-projet CLIME, commune INRIA–ENPC, http://www-rocq.inria.fr/clime/
†Laboratoire CEREA, commun ENPC–EDF R&D, http://cerea.enpc.fr/

La modélisation de la chimie atmosphérique est souvent menée dans un contexte déterministe. La plupart des études et travaux reposent sur un modèle unique, associé à des paramétrisations physiques figées et des données d'entrée invariablement issues des mêmes sources. Cette approche est pourtant mal adaptée à un domaine où de fortes incertitudes entachent toutes les simulations.

1 Introduction à la prévision d'ensemble et à ses objectifs

1.1 Incertitude

1.1.1 Incertitudes de la modélisation

Un modèle de chimie-transport est une représentation de la réalité physique qui repose sur

- l'intégration numérique de l'équation de transport réactif, qui s'écrit pour la concentration c_i d'une espèce i

$$\frac{\partial c_i}{\partial t} + \mathrm{div}\,(V c_i) = \mathrm{div}\left(\rho K \nabla \frac{c_i}{\rho}\right) + \chi_i(c) + S_i - P_i \;; \tag{1}$$

- l'estimation des termes de l'équation 1, à savoir le champ de vent V, la densité de l'air ρ, le tenseur de diffusion K, le terme de chimie $\chi(c)$ et les termes S_i et P_i correspondant respectivement à des sources et des pertes.

En pratique, les termes de l'équation 1 sont soit des données brutes soit, plus souvent, des données estimées par des paramétrisations physiques. Ces paramétrisations physiques reposent elles-mêmes sur des choix de paramètres et des données brutes ou estimées par d'autres paramétrisations.

Il est utile de distinguer trois niveaux dans la modélisation numérique :

1. les données brutes (exemple : les vents issus d'un modèle météorologique) ;
2. les champs paramétrés (exemple : coefficient(s) de diffusion verticale dans le tenseur K) – les paramétrisations associées constituent la formulation du modèle physique ;
3. les schémas numériques pour l'intégration numérique de l'équation 1.

Les valeurs des données brutes sont fortement incertaines : on estime souvent leurs niveaux d'incertitude relative entre 20% et 50% (voir section 2.3). Par exemple, la direction du vent (horizontal) peut être représentée comme une variable aléatoire, de distribution normale, d'écart type de 20° [Hanna *et al.*, 2001]. On considère donc que la direction du vent est connue à ±40° avec un niveau de confiance de 95% – ce qui constitue une incertitude faible en comparaison d'autres champs.

Un champ paramétré peut être estimé par plusieurs paramétrisations physiques concurrentes. Ainsi les vitesses de dépôt d'ozone peuvent être calculées par la paramétrisation issue de Wesely [1989], celle de Zhang *et al.* [2003] ou d'autres. En comparant les vitesses issues de plusieurs paramétrisations, on constate une dispersion d'au moins 30% [Wesely et Hicks, 2000].

Les schémas numériques introduisent de l'erreur et plusieurs schémas numériques peuvent aussi être concurrents. L'intégration numérique introduit donc aussi de l'incertitude.

En conclusion, les trois niveaux constitutifs d'un modèle de chimie-transport sont sujets à des incertitudes souvent élevées. On peut aussi mentionner l'intervention des erreurs de programmation et d'utilisation qui peuvent intervenir dans un modèle. Dans ce contexte, une simulation déterministe, c'est-à-dire reposant sur un modèle unique, a une valeur limitée.

1.1.2 Modèle stochastique

On peut écrire un modèle sous la forme

$$c = \mathcal{M}(p) \tag{2}$$

où les concentrations du vecteur c, indexé par les espèces chimiques, le temps et l'espace, sont estimées par le modèle $\mathcal{M}$ avec les paramètres d'entrée p. Il s'agit d'une formulation déterministe où les entrées du modèle sont prises égales à p et le modèle (formulation physique et schémas numériques) $\mathcal{M}$ est fixé.

On peut réécrire l'équation 2 sous une forme stochastique :

$$\hat{c} = \widehat{\mathcal{M}}(\hat{p}) \tag{3}$$

où les concentrations sont représentées par le vecteur aléatoire $\hat{c}$, les données d'entrée sont introduites par le vecteur aléatoire $\hat{p}$ et le modèle stochastique $\widehat{\mathcal{M}}$ remplace le modèle déterministe.

Les valeurs p des paramètres choisies dans le modèle déterministe sont une réalisation de la variable aléatoire $\hat{p}$, de même que les concentrations c sont une réalisation de la variable aléatoire $\hat{c}$.

Sur cette base on peut distinguer l'erreur d'un modèle et l'incertitude de la modélisation. On note c^t les vraies concentrations. L'*erreur* du modèle générant les concentrations c est une mesure de la distance entre c^t et c, par exemple une norme infinie : $\left\|c - c^t\right\|_\infty$. Une mesure de l'*incertitude* indique la confiance qui peut être accordée *a priori* à une simulation. On suppose que le vecteur aléatoire $\hat{c}$ suit une loi normale $\mathcal{N}(\overline{c}, \Sigma)$ où $\overline{c}$ est l'espérance de $\hat{c}$ et Σ est la matrice de covariance de $\hat{c}$ – c'est-à-dire que $\Sigma = \mathrm{E}\left[(\hat{c} - \overline{c})(\hat{c} - \overline{c})^T\right]$, soit $\Sigma_{i,j} = \mathrm{E}\left[(\hat{c}_i - \overline{c}_i)(\hat{c}_j - \overline{c}_j)\right]$. L'incertitude est précisément décrite par la matrice de covariance Σ.

En pratique, il est difficile de disposer d'une information aussi complète. L'incertitude est alors estimée par un simple scalaire, généralement un écart type ou un écart type relatif (c'est-à-dire l'écart type divisé par la moyenne de la quantité étudiée) d'une quantité scalaire.

1.2 Objectifs

1.2.1 Estimation des incertitudes

Un *ensemble* de modèles échantillonne la distribution des concentrations $\hat{c}$ (équation 3) par des réalisations $\left(c^i\right)_{i \in [\![1,N]\!]}$ où c^i est le vecteur de concentrations du modèle i, et N est le nombre de modèles dans l'ensemble. On note

$$\overline{c}^N = \frac{1}{N} \sum_{i=1}^{N} c^i \tag{4}$$

la moyenne de l'ensemble, appelée *moyenne d'ensemble*. La matrice de covariance de $\hat{c}$ peut être approchée par la variance empirique des N réalisations

$$\Sigma^N = \frac{1}{N-1} \sum_{i=1}^{N} \left(c^i - \overline{c}^N\right) \left(c^i - \overline{c}^N\right)^T \ . \tag{5}$$

À condition d'avoir un ensemble « bien choisi » (se reporter à la section 4), cet ensemble permet donc d'estimer l'incertitude associée à des simulations. Cette information mesure le degré de confiance qui peut être accordé aux simulations, ce qui est un apport important dans le cadre de prises de décision, par exemple.

Cet apport est d'autant plus important que le nombre d'observations est faible en comparaison de la dimension du système. En modélisation de la qualité de l'air, la dimension du système (c'est-à-dire la taille de l'état, ou encore le nombre de concentrations calculées) varie entre 10^5 et 10^7. Le nombre d'observations ne dépasse pas 10^3, et s'approche plus souvent de 10^2... Cela indique que les modèles sont contraints sur peu de composantes. Hors de ces composantes, les résultats sont très incertains.

Cette estimation de l'incertitude est aussi utile en *assimilation de données*. La matrice de covariance d'erreur de l'état, souvent notée B en 4D-Var et P dans le filtre de Kalman, peut être approchée sur la base de l'ensemble. Cette matrice est définie ainsi :

$$P = E\left[(\hat{c} - c^t)(\hat{c} - c^t)^T\right] \tag{6}$$

où c^t est la valeur exacte de l'état (c'est-à-dire des concentrations dans notre cas). Le vecteur exact c^t étant inconnu, on suppose que la moyenne d'ensemble en est une bonne approximation :

$$P \simeq E\left[(\hat{c} - \overline{c}^N)(\hat{c} - \overline{c}^N)^T\right] \ . \tag{7}$$

On obtient naturellement $P \simeq \Sigma^N$. Cette approximation est effectuée dans le filtre de Kalman d'ensemble, qui tire précisément son nom des prévisions d'ensemble qu'il nécessite pour approcher la matrice de covariance d'erreur de l'état.

1.2.2 Prévisions probabilistes

Un système de prévision d'ensemble peut délivrer des probabilités qu'un événement se réalise. Par exemple, en qualité de l'air, la réglementation européenne impose de lancer une alerte si les concentrations d'ozone dépassent 240 $\mu g\,m^{-3}$ (en moyenne horaire). Afin d'anticiper une alerte, il est utile aux autorités de connaître la probabilité qu'un tel événement se réalise. Dans ce cas, la probabilité du dépassement serait par exemple le rapport du nombre de modèles dépassant le seuil sur le nombre total de modèles, ou elle serait calculée sur la base de la moyenne d'ensemble et de la variance de l'ensemble.

De manière plus générale, l'étude des risques passe nécessairement par des estimations de probabilités, et donc potentiellement par des approches d'ensemble. Un exemple critique concerne la dispersion d'un polluant radioactif rejeté à la suite d'un accident nucléaire. Afin d'aider à la prise de décision, un système d'ensemble peut identifier les régions ayant une probabilité significative d'être impactée [Galmarini *et al.*, 2004].

Une description fine de la distribution de probabilité des concentrations est aussi un objectif – de recherche, probablement à long terme – qui permettrait d'obtenir une information plus fine que celle délivrée par certaines méthodes d'assimilation de données.

Par exemple, le filtre de Kalman estime l'espérance des concentrations et leur matrice de covariance d'erreur. Ceci donne aussi l'espérance et la covariance de l'état, soit les deux premiers moments. Une approche d'ensemble permet d'obtenir plus d'informations sur la distribution des concentrations.

1.2.3 Amélioration des prévisions

Un ensemble de modèles apporte évidemment plus d'information qu'un seul modèle, et il est possible d'en tirer parti pour la prévision. Cela peut se faire via l'assimilation de données (filtre de Kalman d'ensemble) ou par combinaison des modèles.

La combinaison de modèles, ou *agrégation de modèles*, consiste à construire une nouvelle prévision par combinaison linéaire des modèles. Les poids de cette combinaison peuvent être calculés par diverses méthodes présentées à la section 5. La combinaison linéaire constitue une nouvelle prévision (et un nouveau modèle) construite pour être plus performante (c'est-à-dire plus proche de la réalité) que les prévisions des autres modèles.

1.3 Formulation stochastique

La grande différence entre la prévision classique et la prévision d'ensemble réside dans l'approche probabiliste de la seconde. Au lieu de simuler une trajectoire possible, on recherche la distribution de probabilité de l'état (des concentrations de polluants). Cette densité de probabilité suit l'équation dite de Fokker-Planck.

1.3.1 Processus de Wiener

Avant d'écrire l'équation de Fokker-Planck, il faut introduire les processus de Wiener. Un processus de Wiener peut être vu comme la limite continue de marches aléatoires. Pour donner l'intuition du comportement d'un processus de Wiener, on peut donner l'exemple de la marche aléatoire $(X_0, X_1, X_2, \ldots)$ où

- $X_0 = 0$;
- $X_{n+1} = X_n - \sqrt{\delta t}$ avec une probabilité $\frac{1}{2}$, ou $X_{n+1} = X_n + \sqrt{\delta t}$ avec une probabilité $\frac{1}{2}$.

On note au passage qu'il s'agit d'une chaîne de Markov puisque $P(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_2 = x_2, X_1 = x_1) = P(X_{n+1} = x | X_n = x_n)$.

Cette marche aléatoire a les propriétés suivantes : son espérance vaut $E(X_n) = 0$ et sa variance vaut $\sigma^2_{X_n} = n\delta t$. La figure 1 représente une suite $(X_n)_n$ possible, pour $\delta t = 1$.

Si on définit $\tilde{X}_t = X_n + (t - n\delta t)(X_{n+1} - X_n)$ pour tout $t \in]n\delta t, (n+1)\delta t]$, alors :

- si $t = n\delta t$, pour tout $a > 0$ et pour tout $s \leq t$, $\tilde{X}_{t+a} - \tilde{X}_t$ est indépendant de $\tilde{X}_s$;
- $\forall t \quad E(\tilde{X}_t) = 0$;
- $\forall t \quad E(\tilde{X}_t^2) = t$.

Le processus $\tilde{X}_t$ a des propriétés similaires à un processus de Wiener.

Un processus de Wiener, noté W_0, est défini comme un processus stochastique ayant les propriétés suivantes :

- $W_0 = 0$;
- pour tout $t \geq 0$ et tout $0 \leq s < t$, l'incrément $W_t - W_s$ suit une loi normale de moyenne 0 et de variance $t - s$;

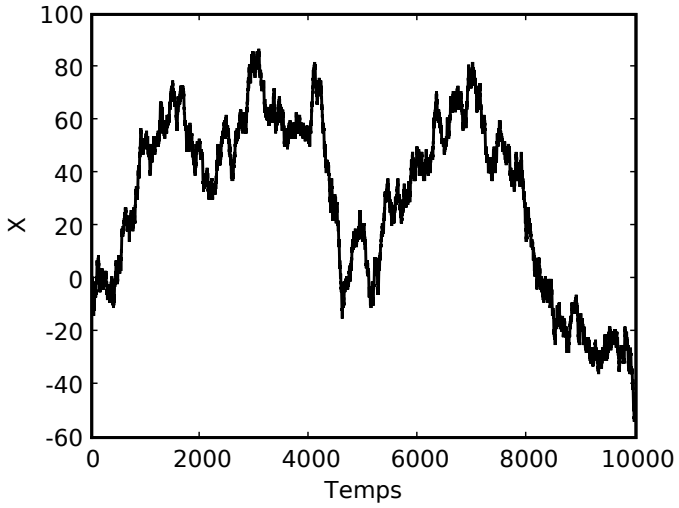


FIG. 1 – Exemple de marche aléatoire. La suite $(X_n)_n$ est définie par $X_0 = 0$ et $X_{n+1} = X_n - 1$ avec une probabilité $\frac{1}{2}$, ou $X_{n+1} = X_n + 1$ avec une probabilité $\frac{1}{2}$.

– pour tous t_1, s_1, t_2 et s_2 tels que $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$, les incréments $W_{t_1} - W_{s_1}$ et $W_{t_2} - W_{s_2}$ sont indépendants.

1.3.2 Équation de Fokker-Planck

Considérons un système déterministe régi par l'équation aux dérivées ordinaires

$$\mathrm{d}x_t = f(x_t, t)\mathrm{d}t \ , \tag{8}$$

où x_t et $f(x_t, t)$ sont des vecteurs de dimension n.

Si ce système modélise un phénomène partiellement connu, le modèle exact diffère du modèle déterministe d'un terme aléatoire :

$$\mathrm{d}x_t = f(x_t, t)\mathrm{d}t + g(x_t, t)\mathrm{d}w_t \ , \tag{9}$$

où w_t est un vecteur de dimension m dont chaque composante est un processus de Wiener, et $g(x_t, t)$ est une matrice de dimension $n \times m$. La notation $\mathrm{d}w_t$ est introduite sans plus de détails. On se reportera au calcul d'Itô pour préciser le cadre mathématique.

L'équation 9 est le modèle stochastique qui décrit l'évolution temporelle de la variable aléatoire x_t. L'équation de Fokker-Planck décrit l'évolution temporelle de la densité de probabilité $P(x, t)$ associée à x_t :

$$\frac{\partial P}{\partial t} = -\operatorname{div}(f(x,t)P) + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[Q_{i,j}P\right] \tag{10}$$

où Q est une matrice de taille $n \times n$ définie par $Q(x, t) = g(x, t)g(x, t)^T$.

Le membre de droite est composé de deux termes :

– un terme d'advection $\operatorname{div}(f(x, t)P)$ qui décrit la propagation de la densité de probabilité sous l'effet du modèle déterministe ;

– un terme de diffusion $\frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[Q_{i,j}P\right]$ qui décrit les effets de la composante aléatoire.

Prenons l'exemple de l'équation de Langevin (dimension 1) :

$$\mathrm{d}x_t = -x_t \mathrm{d}t + \sqrt{q}\mathrm{d}w_t \tag{11}$$

où q ne dépend ni de x_t ni de t.

L'équation de Fokker-Planck associée est

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x}(xP) + \frac{q}{2}\frac{\partial^2 P}{\partial x^2}\ . \tag{12}$$

Approcher numériquement la densité de probabilité demande de résoudre une équation d'advection-diffusion dans l'espace des valeurs potentiellement prises par l'état x. Le coût de la résolution numérique de l'équation est très nettement supérieur à celui de l'équation scalaire du modèle déterministe ($\frac{\mathrm{d}x_t}{\mathrm{d}t} = -x_t$). Dans le cas multidimensionnel (par exemple $n = 10^5$ ou $n = 10^7$, ce qui est courant en qualité de l'air), le coût d'une résolution numérique de l'équation de Fokker-Planck est prohibitif.

Il vient de plus s'ajouter des contraintes spécifiques à l'intégration en temps d'une densité de probabilité : conservation de l'intégrale de la densité (qui doit valoir 1) et contrainte de positivité.

Notons enfin le cas particulier où $f(x_t, t) = 0$ et $g(x_t, t) = \mathrm{cte}$ qui correspond au cas de diffusion pure. Par exemple, la densité de probabilité $P(t, x)$ de présence au temps t et à la position x d'une particule ayant un mouvement brownien suit l'équation de diffusion $\frac{\partial P}{\partial t} = \mathrm{cte}\frac{\partial^2 P}{\partial x^2}$.

2 Simulations Monte Carlo

Afin d'approcher la distribution de probabilité des concentrations calculées par un modèle de chimie-transport, ou simplement un paramètre de cette distribution (moyenne, variance), les méthodes de Monte Carlo sont couramment utilisées.

2.1 Principe

Si on souhaite estimer la distribution de la variable aléatoire $\widehat{\mathcal{M}}(X)$ ($\widehat{\mathcal{M}}$ est un modèle stochastique) sur la base de la distribution de la variable aléatoire X, les méthodes de Monte Carlo proposent d'effectuer de multiples tirages dans l'espace des états possibles (de X). Ces tirages peuvent être aléatoires, pseudo-aléatoires (en pratique) ou déterministes. En augmentant leur nombre, l'espace des états devient mieux échantillonné, et la distribution de $\widehat{\mathcal{M}}(X)$ s'affine.

Les méthodes de Monte Carlo sont souvent illustrées dans un cadre non stochastique où il s'agit d'estimer la valeur de l'intégrale I d'une fonction f :

$$I = \int_{[0,1]^s} f(x)dx \tag{13}$$

où s est la dimension du vecteur d'état x, et f est une fonction à valeur dans $\mathbb{R}$.

Plusieurs tirages aléatoires $X_1, X_2, \ldots, X_N$ dans $[0,1]^s$ sont effectués indépendamment et uniformément. L'intégrale I est alors approchée par

$$I_N = \frac{1}{N} \sum_{i=1}^{N} f(X_i) \ . \tag{14}$$

Les tirages sont indépendants et suivent tous la même loi de probabilité (uniforme). Il en est donc de même pour les $f(X_i)$. Il est possible d'appliquer le théorème central limite. On note $\sigma^2 = \int_{[0,1]^s} (f(x) - I)^2 dx$ la variance de f. Selon le théorème central limite, la suite

$$U_N = \sqrt{N} \frac{I_N - I}{\sigma} \tag{15}$$

converge en loi vers la loi normale centrée réduite $\mathcal{N}(0,1)$.

Ceci prouve que l'approximation I_N estime I avec un intervalle de confiance dimensionné par $\frac{\sigma}{\sqrt{N}}$. Il s'agit d'un résultat fondamental pour les simulations Monte Carlo : la convergence est assez lente (en $\sqrt{N}$), dépendante de la variance de la fonction f, et indépendante de la dimension s.

Le même principe peut être appliqué pour approcher la distribution de $\widehat{\mathcal{M}}(X)$ (mentionné précédemment), ou un moment de sa distribution.

2.2 Techniques et méthodes associées

2.2.1 Méthodes de quasi Monte Carlo

Les méthodes de Monte Carlo reposent sur la génération de nombres aléatoires ou pseudo-aléatoires. Il n'y a alors pas de garantie que les échantillons soient répartis avantageusement dans l'espace des états possibles. Les méthodes de quasi Monte Carlo échantillonnent de manière déterministe la variable aléatoire d'entrée X. Leur objectif est d'assurer une répartition des tirages représentative de la distribution (de X).

2.2.2 Méthodes de réduction de variance

Comme montré à la section 2.1, la vitesse de convergence d'une méthode Monte Carlo dépend de la variance de la fonction cible f. Dans le cas du calcul d'une intégrale, on peut chercher à réduire cette variance en écrivant

$$I = \int_{[0,1]^s} g(x)dx + \int_{[0,1]^s} (f(x) - g(x))dx \ , \tag{16}$$

où g est une fonction connue, facilement intégrable, et de sorte que la variance de $f-g$ soit plus faible que celle de f. Le premier terme $\int_{[0,1]^s} g(x)dx$ est connu. Il reste à déterminer $\int_{[0,1]^s} (f(x) - g(x))dx$. Puisque la nouvelle fonction à intégrer, $f - g$, est de variance inférieure à celle de f, les simulations Monte Carlo convergeront plus vite que dans le cas de f.

Une variante est la méthode d'échantillonnage préférentiel (en anglais, « importance sampling ») où l'idée est d'échantillonner surtout les régions importantes. Dans le calcul d'une intégrale, les régions où f est faible sont peu échantillonnées car elles contribuent peu à la valeur finale et donc à l'erreur finale.

Une autre méthode consiste à diviser l'intervalle d'intégration en n intervalles sur lesquels une méthode de Monte Carlo classique est appliquée. Cette méthode dite d'échantillonnage stratifié (en anglais, « stratified sampling ») produit l'approximation suivante :

$$\tilde{I} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{1}{n_i} f(x_{i-1} + (x_i - x_{i-1})\alpha_{i,j}) \tag{17}$$

où $0 = x_0 < x_1 < \ldots < x_n = 1$ définissent les bornes des sous-intervalles et les $\alpha_{i,j}$ sont tirés indépendamment et uniformément dans $[0, 1]$. La variance de f dans l'intervalle i vaut

$$\frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x^i} f^2(x)\mathrm{d}x - \left(\frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x^i} f(x)\mathrm{d}x \right)^2 . \tag{18}$$

Le nombre d'échantillons dans chaque intervalle peut être choisi en conséquence.

2.3 Propagation des incertitudes sur les champs d'entrée

Les méthodes de Monte Carlo peuvent s'appliquer aux incertitudes ayant leur source dans les données d'entrée continues. Les données d'entrées sont soit les coefficients de l'équation d'advection-diffusion-réaction 1, soit des données nécessaires aux paramétrisations physiques (voir section 1.1.1).

On suppose que le modèle $\mathcal{M}$ est déterministe et on pose

$$\hat{c} = \mathcal{M}(\hat{p}) \tag{19}$$

où le vecteur aléatoire $\hat{p}$ représente l'ensemble des données d'entrée continues et incertaines, et $\hat{c}$ est la variable aléatoire représentant les concentrations.

2.3.1 Modélisation des incertitudes

L'étape essentielle consiste à associer aux éléments $\hat{p}$ une loi de probabilité selon laquelle les échantillons de la méthode Monte Carlo sont tirés. Cette étape de modélisation des incertitudes est cruciale car elle conditionne directement les incertitudes sur les concentrations calculées.

Une indication sur la validité de la modélisation peut être faite *a posteriori*, en estimant la qualité de l'ensemble sur la base des observations (voir section 4). En l'absence d'observations, les incertitudes en entrée peuvent éventuellement être majorées ou minorées selon l'objectif des simulations. Par exemple, l'évaluation d'un risque maximal conduit naturellement à se placer dans les hypothèses les plus défavorables.

On considère généralement qu'un paramètre, une donnée, suit une loi lognormale ou, parfois, normale – voir annexe A. La loi lognormale est préférentiellement utilisée pour

plusieurs raisons. D'abord, elle apparaît souvent naturellement dans les données environnementales. Ensuite, d'après le théorème central limite, elle est la limite d'un produit de variables indépendantes et identiquement distribuées. De plus, la multiplication ou la division de variables de loi lognormale suit une loi lognormale. Enfin, beaucoup de variables physique ont une contrainte de positivité que la loi lognormale respecte.

En pratique, on utilise une version légèrement différente : la loi est tronquée. Les valeurs très éloignées de la médiane sont supprimées – la densité de probabilité est annulée. Ceci évite de considérer des valeurs physiquement irréalistes.

Les paramètres de la loi sont choisis sur les recommandations d'experts, c'est-à-dire de ceux qui fournissent les données d'entrée (météorologues, chimistes, etc.). Ces derniers se fondent sur (1) les comparaisons entre leurs modèles et les données d'observations, (2) la variabilité entre différents modèles, (3) parfois sur des prévisions d'ensemble (pour les domaines les plus avancés).

Des exemples d'incertitudes, issues de Hanna *et al.* [2001], sont présentés dans le tableau 1 pour des simulations continentales. L'intervalle de confiance à 95% de plusieurs champs (de distribution lognormale avec $\alpha = 2$) est compris entre la moitié de la valeur et le double de la valeur du champ.

TAB. 1 – Incertitudes associées à quelques champs d'entrée d'un modèle de chimie-transport à une échelle continentale. Pour chaque champ, l'incertitude d'un paramètre $\hat{p}$ est mesurée avec un intervalle de confiance à 95%. Pour la loi lognormale, cet intervalle est défini par un facteur α tel que $\hat{p}$ soit dans l'intervalle $[\frac{p}{\alpha}, \alpha p]$ avec une probabilité de 95% et où p est la valeur médiane de $\hat{p}$. Les valeurs sont extraites de Hanna *et al.* [2001].

Champ	Loi	Incertitude
Conditions aux limites supérieures d'ozone	lognormale	$\alpha = 1.5$
Conditions aux limites supérieures de NO_x	lognormale	$\alpha = 3$
Conditions aux limites latérales d'ozone	lognormale	$\alpha = 1.5$
Conditions aux limites latérales de NO_x	lognormale	$\alpha = 3$
Points d'émission de NO_x majeurs	lognormale	$\alpha = 1.5$
Vitesse du vent	lognormale	$\alpha = 1.5$
Direction du vent	normale	$\pm 40°$
Température	normale	± 3 K
Diffusion verticale (nuit)	lognormale	$\alpha = 3$
Précipitations	lognormale	$\alpha = 2$
Eau liquide	lognormale	$\alpha = 2$
Émissions biogéniques	lognormale	$\alpha = 2$
Constantes photolytiques	lognormale	$\alpha = 2$

Les incertitudes doivent être ajustées selon l'échelle de simulation (taille du domaine). En particulier, plus des données sont moyennées temporellement ou spatialement moins elles sont incertaines.

Les lois et incertitudes sont introduites pour des champs bidimensionnels ou tridimen-

sionnels (en espace) et dépendant du temps. En pratique, pour un champ $\hat{p}$ de valeur médiane p dépendant du temps t et de la position spatiale x, une perturbation est appliquée à l'ensemble du champ de sorte que chaque variable $\hat{p}(t,x)$ suive la loi prescrite. Par exemple, pour une loi lognormale, on écrit

$$\forall t, x \quad \hat{p}(t,x) = p(t,x) \times \sqrt{\alpha}^{\gamma} \ , \tag{20}$$

où γ suit une loi normale centrée réduite, indépendante de t et x. Un unique tirage de γ est effectué pour toutes les valeurs d'un échantillon de $\hat{p}$.

Un point délicat concerne les corrélations temporelles et spatiales entre les différentes valeurs du champ. Avec la perturbation appliquée à l'équation 20, on vérifie facilement que la corrélation entre deux valeurs $\ln \hat{p}(t_0, x_0)$ et $\ln \hat{p}(t_1, x_1)$ vaut 1. Or les sources d'incertitudes en deux points éloignés ne sont pas les mêmes. Une modélisation fine de l'incertitude conduirait à faire dépendre γ du temps t et de la position x.

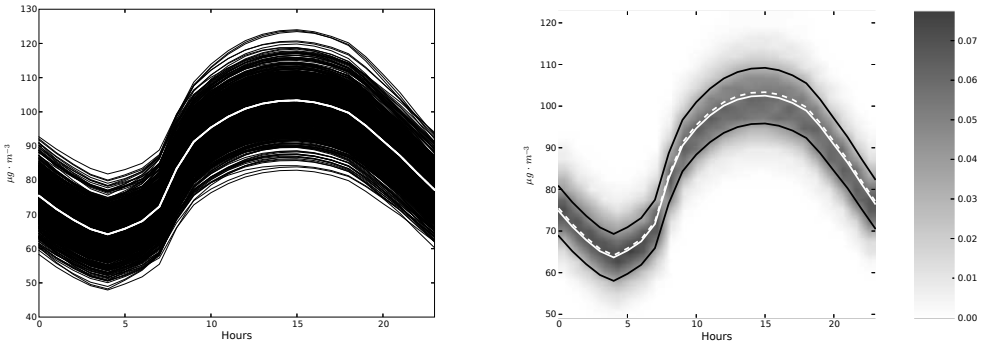
2.3.2 Exemple de simulations Monte Carlo

Des exemples de simulations Monte Carlo appliquées à la qualité de l'air sont exposés dans Hanna *et al.* [1998]; Beekmann et Derognat [2003] (efficacité de réductions d'émissions) et Hanna *et al.* [2001]; Hanna et Davis [2002] (estimation de l'incertitude des concentrations simulées). Dans Mallet [2005], les données sont perturbées selon la procédure précédemment décrite, avec les incertitudes du tableau 2. Les incertitudes sont minorées de sorte que les simulations Monte Carlo conduisent à une dispersion minimale des polluants. Il s'agit d'obtenir un minorant de l'incertitude. La simulation couvre une partie de l'Europe et s'étend sur une semaine de l'été 2001. Huit cents simulations sont effectuées.

Les profils journaliers moyens d'ozone sont représentés à la figure 2. Pour chaque heure de la journée, la valeur du profil d'une simulation est la moyenne des concentrations calculées sur tout le domaine et tous les jours de la semaine simulée. Il s'agit donc de valeurs fortement moyennées et qui présentent pourtant une grande dispersion. La densité de probabilité est estimée sur la base de l'ensemble de profils.

Tab. 2 – Incertitude sur les données d'entrée [Mallet, 2005]. Les entrées ont la même signification que dans le tableau 1. Les incertitudes sont minorées de sorte que les simulations Monte Carlo permettent d'estimer une incertitude minimale sur les concentrations de sortie.

Champ	Loi	Incertitude
Atténuation nuageuse	lognormale	1.3
Vitesses de dépôt (O_3 et NO_2)	lognormale	1.3
Conditions aux limites (O_3)	lognormale	1.2
Émissions anthropogéniques	lognormale	1.5
Émissions biogéniques	lognormale	2.0
Constantes photolytiques	lognormale	1.3



(a) Ensemble des 800 profils (b) Densité de probabilité

FIG. 2 – Profils journaliers moyens d'ozone. Pour chaque heure de la journée, la valeur du profil d'une simulation est la moyenne des concentrations calculées sur tout le domaine et tous les jours de la semaine simulée. Dans (a), les 800 profils sont représentés. En blanc (et au centre), on représente la simulation de référence (c'est-à-dire sans perturbations). Dans (b), la densité de probabilité (dégradés) est représentée avec l'espérance (courbe continue en blanc), l'espérance à laquelle l'écart type est soustrait ou ajouté (courbes en noir) et le profil de la simulation de référence (courbe discontinue en blanc).

Afin de s'assurer de la validité des résultats, il est important de vérifier que la procédure Monte Carlo a convergé. Pour cela, on étudie le comportement des quantités-cibles en fonction du nombre de simulations. La figure 3 illustre la méthode.

La figure 4 montre la répartition spatiale de l'écart type de concentrations d'ozone. Elle illustre la richesse des sorties de la procédure. Le coût de telles simulations est certes élevé (ici, 800 fois une simulation classique), mais elle permet d'accéder à une information fine sur les incertitudes liées aux données d'entrée.

3 Simulations multi-modèles

À la section 1.1.1, la deuxième source d'incertitude identifiée est la formulation du modèle physique, c'est-à-dire les paramétrisations physiques utilisées pour estimer les termes de l'équation de dispersion réactive 1. Cette incertitude ne peut être représentée avec précision par des simulations Monte Carlo. En effet, les changements dans la formulation du modèle sont discrets. On a alors recours à des simulations multi-modèles. On parle aussi de prévision d'ensemble.

Notons que les données d'entrée peuvent aussi être perturbées s'il s'agit de constituer un ensemble représentatif de toute l'incertitude.

3.1 Construction des modèles

Le principe de génération d'un ensemble est simple. Dans la formulation du modèle, tous les champs estimés sur la base de paramétrisations physiques sont considérés. Pour chaque champ, toutes les paramétrisations acceptables sont recensées. Construire un modèle consiste à sélectionner une paramétrisation pour chaque champ à estimer. En choisissant

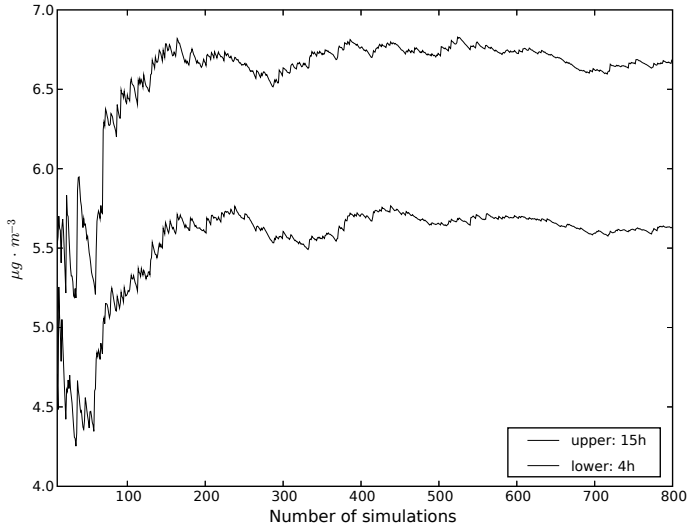


FIG. 3 – Écart type de la moyenne spatio-temporelle des concentrations d'ozone à 4 h et à 15 h sur les sept derniers jours simulés, en fonction du nombre de simulations. À partir de 200 simulations, voire 400 simulations (selon l'exigence), on peut considérer que la convergence s'est effectuée.

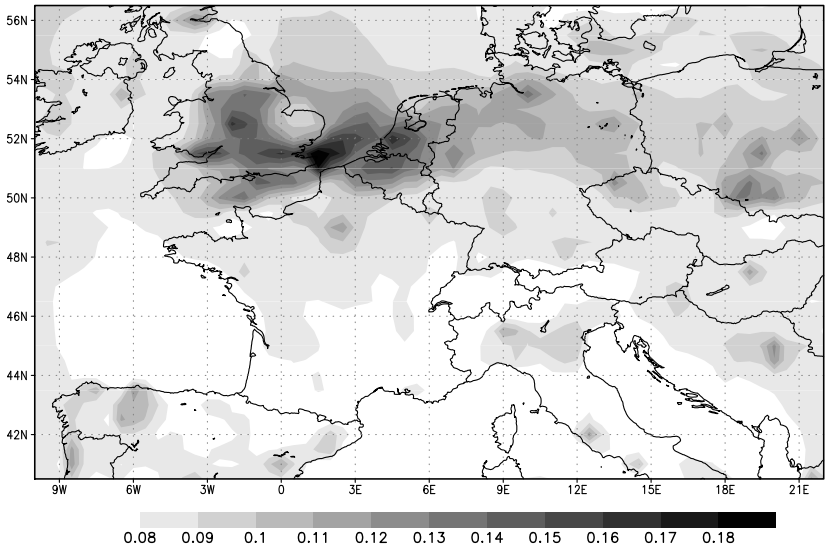


FIG. 4 – Répartition spatiale de l'incertitude relative (la moyenne temporelle de l'écart type divisée par la concentration moyenne).

d'autres paramétrisations, on construit d'autres modèles, et finalement un ensemble de modèles.

Un point-clé réside évidemment dans le choix des paramétrisations. Pour chaque champ, il faut d'abord constituer la liste de toutes les paramétrisations intéressantes pour l'estimer. Il peut s'agir de paramétrisations concurrentes, c'est-à-dire de valeurs égales, ou de paramétrisations dégradées, c'est-à-dire moins performantes que la ou les paramétrisations de référence. Les paramétrisations dégradées peuvent néanmoins apporter de l'information. Localement, elles peuvent se comporter de manière plus satisfaisante que la ou les paramétrisations de référence. Elles doivent donc être considérées dans la constitution d'un ensemble.

Une erreur commune consiste à objecter qu'une paramétrisation inférieure (dans le sens où elle estime moins bien les véritables valeurs d'un champ) est « fausse ». Cette appréciation marque l'empreinte d'une approche déterministe, qui ne reconnaît souvent que les bonnes paramétrisations et les mauvaises paramétrisations. Or toute paramétrisation est fausse puisqu'elle ne délivre qu'une approximation de la réalité. L'approche stochastique invite à classer les paramétrisations selon leur *probabilité*. Une paramétrisation est plus ou moins probable. Une paramétrisation qu'on serait tenter de qualifier de fausse est en fait peu probable. Elle doit pourtant être incluse dans un ensemble de simulation car sa probabilité n'est pas nulle. Elle doit toutefois être moins représentée dans l'ensemble qu'une paramétrisation de plus grande valeur (plus probable).

De même, un ensemble de modèles ne peut pas être constitué que de modèles jugés bons selon les critères déterministes. Ces derniers modèles sont réglés de sorte à s'accorder au mieux avec les observations. Leurs performances sont mesurées, par exemple, par une erreur quadratique moyenne. Du fait de leurs bonnes performances, ces modèles sont situés dans des zones où la densité de probabilité est forte. Construire un ensemble représentant correctement l'incertitude suppose d'inclure des modèles peu probables, donc situés en des points de densité de probabilité faible.

Enfin, les membres de l'ensemble doivent rarement être choisis aléatoirement ou indépendamment. Dans le cas de simulations Monte Carlo (section 2.3.2), un minimum de 200 échantillons est nécessaire à la convergence de l'écart type (mesure de l'incertitude) de l'ensemble. La variance due à l'incertitude dans la formulation du modèle est généralement du même ordre que celle due aux incertitudes dans les données d'entrée. Par conséquent, le nombre d'échantillons nécessaires à l'estimation de l'incertitude est du même ordre. Or, à cause de contraintes technologiques, il est difficile de réaliser autant de tirage, de surcroît aléatoirement, lorsque les changements affectent la formulation du modèle. Il faut donc effectuer des tirages contrôlés, censés échantillonner correctement la distribution des concentrations avec peu de modèles (quelques dizaines).

3.2 Exemple : étude des incertitudes en prévision des concentrations d'ozone

On poursuit sur l'exemple de la prévision photochimique sur l'Europe. Parmi les composantes prépondérantes de la formulation du modèle, on compte le mécanisme chimique (cycle de l'ozone) et la paramétrisation pour la diffusion verticale.

Plusieurs mécanismes chimiques existent pour modéliser le cycle de l'ozone. Le nombre d'espèces chimiques dans le mécanisme peut varier (et donc la taille du vecteur d'état).

Les réactions prises en compte diffèrent aussi. Une même réaction peut aussi avoir des taux de réaction différents d'un mécanisme à l'autre. Les mécanismes les plus connus sont CBM IV [Gery *et al.*, 1989], RADM [Stockwell *et al.*, 1990] et RACM [Stockwell *et al.*, 1997]. On peut se référer à Gross et Stockwell [2003] pour une comparaison entre trois mécanismes dont RADM et RACM.

La paramétrisation qui estime le coefficient de diffusion verticale peut être celle issue de Troen et Mahrt [1986] ou celle de Louis [1979]. La première est considérée comme plus adaptée à un modèle de chimie-transport. En effet, elle ne dépend pas de la résolution du modèle et un modèle de chimie-transport repose généralement sur une discrétisation verticale grossière. Au contraire, la paramétrisation de Louis requiert l'évaluation de gradients horizontaux de vent, ce qui peut être source d'erreurs importantes.

En exemple, le tableau 3 liste des paramétrisations physiques, des choix de données d'entrée et des choix numériques qui permettent de constituer un ensemble de modèle.

Tab. 3 – Exemple de paramétrisations, choix de données d'entrée et de choix numérique qui peuvent intervenir dans la construction d'un modèle. Extrait de Mallet [2005].

n°	Modèle	Référence	Alternative	Commentaire
		Paramétrisations physiques		
1.	Chimie	RACM	RADM 2 [Stockwell *et al.*, 1990]	
2.	Diffusion verticale	Troen & Mahrt	Louis [Louis, 1979]	
3.			Louis pour les conditions stables	Troen & Mahrt conservé en conditions instables
4.	Vitesses de dépôt	Zhang [Zhang *et al.*, 2003]	Wesely [Wesely, 1989]	
5.	Flux de surface	Flux de chaleur	Flux de moment	Utilisé dans le calcul de la résistance aérodynamique (vitesses de dépôt)
6.	Atténuation nuageuse	méthode RADM [Chang *et al.*, 1987; Madronich, 1987]	Esquif (ESQUIF [2001])	
7.	Humidité relative critique	Fonction de σ	Constante sur deux niveaux	Utilisée dans le calcul de l'atténuation nuageuse dans la méthode RADM
		Données d'entrée brutes		
8.	Distribution verticale des émissions	Toutes dans la première couche	Toutes dans les deux premières couches	
9.	Occupation des sols	USGS	GLCF	Pour le calcul des vitesses de dépôt
10.	Occupation des sols	USGS	GLCF	Pour le calcul des émissions biogéniques
11.	Exposant p dans Troen & Mahrt	2	3	
12.	Constantes photolytiques	JPROC (de Models-3, EPA)	Fonction de l'angle zénithal	
		Approximations numériques		
13.	Pas de temps	600 s	100 s	
14.			1800 s	
15.	Résolution verticale	5 niveaux	9 niveaux	La hauteur de la première couche demeure 50 m
16.	Hauteur de la première couche	50 m	40 m	La hauteur supérieure des autres couches ne change pas
17.	Équation de continuité	$\mathrm{div}\,(V) = 0$	$\mathrm{div}\,(\rho V) = 0$	
		Autres données perturbées		
18.	Hauteur de couche limite	ECMWF	Augmentée de 10%	
19.	Émissions de NO	EMEP	Augmentées de 25%	Émissions biogéniques incluses
20.	Émissions biogéniques	Simpson *et al.* [1999]	Augmentées de 100%	À l'exclusion des émissions biogéniques de NO
21.	Conditions aux limites d'ozone	Mozart 2	Diminuées de 10%	

Les profils journaliers moyens d'ozone, pour les 48 membres (modèles) d'un ensemble, sont représentés à la figure 5. La figure 6 montre une carte d'incertitude relative. Des détails sont disponibles dans Mallet [2005].

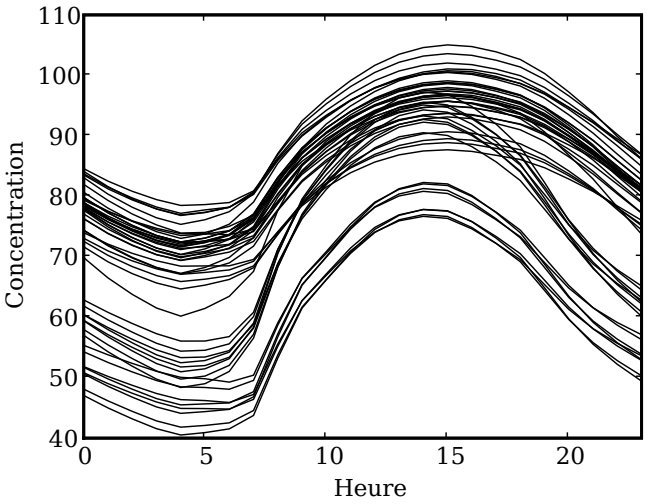


FIG. 5 – Profils journaliers moyens d'ozone ($\mu g\ m^{-3}$) pour 48 modèles d'un ensemble. Pour chaque heure de la journée, la valeur du profil d'une simulation est la moyenne des concentrations calculées sur tout le domaine (européen) et tous les jours des quatre mois simulés.

4 Évaluation de la qualité d'un ensemble

On suppose qu'on dispose d'un ensemble $\mathcal{E}$ de modèles M^m pour $m \in [\![1, N]\!]$ où N est le cardinal de $\mathcal{E}$, donc le nombre de modèles.

4.1 Qu'est-ce qu'un bon ensemble ?

Les objectifs de la prévision d'ensemble sont exposés à la section 1.2. Un bon ensemble permet de remplir un ou plusieurs objectifs :

1. permettre une estimation de l'incertitude, avoir une dispersion représentative de l'incertitude, c'est-à-dire, en pratique, avoir un écart type empirique sur des quantités-cibles conforme à l'écart type de la variable aléatoire correspondante ;
2. estimer au mieux les probabilités d'un événement ;
3. être propice à des combinaisons linéaires de modèles performantes en prévision.

4.2 Indicateurs de la valeur d'un ensemble

Il existe plusieurs indicateurs de la qualité d'un ensemble. Tous reposent sur des comparaisons aux observations puisque ces dernières sont la seule information disponible hors des modèles.

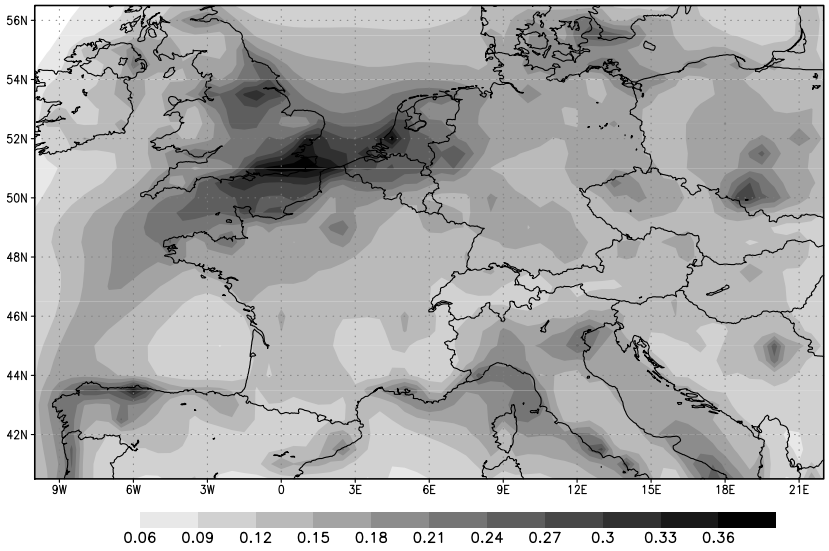


FIG. 6 – Distribution spatiale de la dispersion d'un ensemble de 48 modèles. L'écart type de l'ensemble est calculé dans chaque cellule et à chaque heure. Ensuite, ces écarts types sont moyennés (en temps) dans chaque cellule et divisés par la concentration moyenne de la cellule, ce qui rend un écart type relatif. On constate des incertitudes élevées le long des côtes.

Par exemple, si les observations d'une station sont ajoutées à l'ensemble (c'est-à-dire aux prévisions de tous les modèles au point d'observation), la série temporelle des observations ne doit pas avoir de caractéristiques qui la particularisent parmi les séries temporelles des modèles. En particulier, les observations doivent donc avoir un niveau et une variabilité similaires à plusieurs modèles de l'ensemble.

4.2.1 Diagramme de rang

Le diagramme de rang, aussi appelé diagramme de Talagrand, représente la répartition des observations vis-à-vis des modèles de l'ensemble. Pour chaque observation, les concentrations des modèles sont classées de la plus petite à la plus grande. La position de l'observation par rapport aux modèles est repérée par un indice : 0 si l'observation est en dessous de tous les modèles, 1 si elle est située entre la concentration la plus faible et la concentration immédiatement supérieure, ..., N si l'observation est supérieure à toutes les concentrations simulées. L'opération est répétée pour toutes les observations et le nombre d'occurences de chaque indice est calculé. Ainsi, par exemple, le nombre d'observations associées à l'indice 0 est le nombre d'observations qui sont inférieures à tous les modèles. Le nombre d'occurence est représenté par un diagramme fonction de l'indice. La figure 7 en donne un exemple.

Un diagramme en forme de « U », c'est-à-dire où un grand nombre d'observations sont soit en dessous de l'enveloppe inférieure de l'ensemble soit au-dessus de l'enveloppe supérieure, dénote un manque de dispersion de l'ensemble. L'incertitude est sous-estimée. Un diagramme avec quelques valeurs fortes au centre dénote au contraire une trop grande dispersion. Par ailleurs, un manque de symétrie dans le diagramme indique un biais

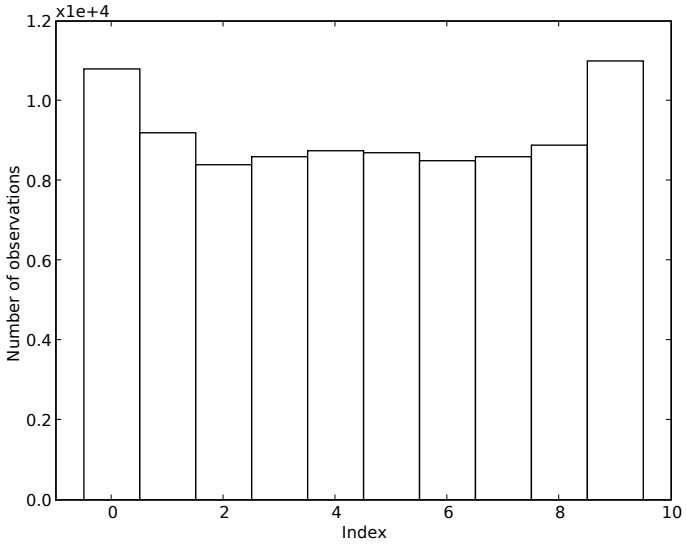


FIG. 7 – Diagramme de rang (pour neuf modèles) : nombre d'observations situées entre chaque modèle.

systématique. On préfère donc un diagramme relativement uniforme.

Un diagramme parfaitement uniforme n'est pas nécessairement un bon diagramme. On suppose qu'un ensemble échantillonne une loi normale $\mathcal{N}(\mu, \sigma^2)$ avec trois modèles. Ces modèles sont choisis de sorte à ce que leurs sorties prennent les valeurs $\mu - \alpha\sigma$, μ et $\mu + \alpha\sigma$, avec α à déterminer. Un diagramme de rang uniforme s'obtient si $\mu - \alpha\sigma$ est le 25e percentile et $\mu + \alpha\sigma$ le 75e percentile. En effet, μ est le 50e percentile, donc 25% des observations (supposées exactes) tombent dans $]-\infty, \mu - \alpha\sigma]$, 25% dans $[\mu - \alpha\sigma, \mu]$, 25% dans $[\mu, \mu + \alpha\sigma]$ et les derniers 25% dans $[\mu + \alpha\sigma, +\infty[$. Le diagramme de rang est alors parfaitement uniforme. Dans ce cas, il faut choisir $\alpha \simeq 0.6745$. Cependant, si l'ensemble doit être représentatif de l'incertitude, on peut souhaiter que son écart type empirique soit égal à σ, c'est-à-dire

$$\frac{1}{2}\left((\mu - \alpha\sigma - \mu)^2 + (\mu - \mu)^2 + (\mu + \alpha\sigma - \mu)^2\right) = \sigma^2, \tag{21}$$

ce qui est vérifié pour $\alpha = 1$. Un diagramme uniforme n'est donc pas souhaitable.

4.2.2 Score de Brier

Le score de Brier est à voir comme le pendant probabiliste de l'erreur quadratique moyenne (très utilisée pour analyser les performances des modèles déterministes). Un score de Brier se calcule pour un événement particulier, par exemple, l'événement $c \geq 240\ \mu\mathrm{g\,m^{-3}}$ où c est la concentration d'un polluant. À chaque échéance de prévision i, la probabilité selon l'ensemble que cet événement se réalise est notée p_i. Elle peut être estimée avec le nombre de modèles réalisant l'événement :

$$p_i = \frac{|\{m/M_i^m \geq 240\ \mu\mathrm{g\,m^{-3}}\}|}{N}\ . \tag{22}$$

La probabilité exacte (observée) que l'événement se réalise vaut soit 1 si l'événement se réalise effectivement, soit 0 si l'événement ne se réalise pas. On note o_i cette probabilité déterminée par l'observation.

Le score de Brier se calcule alors sur la base de N échéances observées :

$$B = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2 \ . \tag{23}$$

Il peut être intéressant de faire le rapport entre le score de Brier d'un ensemble et le score de Brier des modèles pris individuellement. Pour un modèle pris individuellement, la probabilité sera prise égale à 1 ou 0 selon que le modèle prédit la réalisation ou non de l'événement.

4.2.3 Diagramme de fiabilité

Pour événement particulier, le diagramme de fiabilité représente la fréquence d'apparition d'une certaine probabilité dans l'ensemble et dans les observations.

Pour construire ce diagramme, l'intervalle des probabilités $[0, 1]$ est découpé en sous-intervalles. On suppose que le premier intervalle est $[0, 0.1]$. On sélectionne ensuite l'ensemble des échéances pour lesquelles la probabilité estimée (sur la base de l'ensemble) de l'événement est dans l'intervalle $[0, 0.1]$. On calcule ensuite la probabilité moyenne exacte de l'événement parmi ces échéances.

Par exemple, pour l'événement $c \geq 240\ \mu g\, m^{-3}$, on constitue d'abord l'ensemble des échéances

$$\mathcal{I}_{[0,0.1]} = \{i / p_i \in [0, 0.1]\} \ , \tag{24}$$

où une échéance observée est repérée par l'indice i et p_i est la probabilité de réalisation de l'événement, par exemple calculée selon la formule 22. Pour ces échéances, la probabilité moyenne exacte de réalisation est

$$Y_{[0,0.1]} = \frac{\sum_{i \in \mathcal{I}_{[0,0.1]}} o_i}{|\mathcal{I}_{[0,0.1]}|} \ , \tag{25}$$

où $|\mathcal{I}_{[0,0.1]}|$ est le cardinal de $\mathcal{I}_{[0,0.1]}$.

L'opération est répétée pour tous les sous-intervalles. Le diagramme de fiabilité consiste à représenter $\left(Y_{[x_i,x_{i+1}]}\right)_i$ en fonction de $\left(\frac{x_i+x_{i+1}}{2}\right)_i$.

Un diagramme de fiabilité est donné en exemple à la figure 8. Un diagramme parfait est tel que $Y_{[x_i,x_{i+1}]} = \frac{x_i+x_{i+1}}{2}$. Un point en dessous de la première diagonale correspond à surestimation de la probabilité, et vice-versa.

5 Agrégation de prévisions

L'agrégation de prévisions consiste à combiner (linéairement) les prévisions de plusieurs modèles de sorte à construire une nouvelle prévision. Le poids associé à chaque modèle est choisi pour que la combinaison linéaire soit plus performante que chaque modèle pris individuellement. Par « performant », on entend proche des observations selon une mesure donnée. La mesure en question est souvent l'erreur quadratique moyenne.

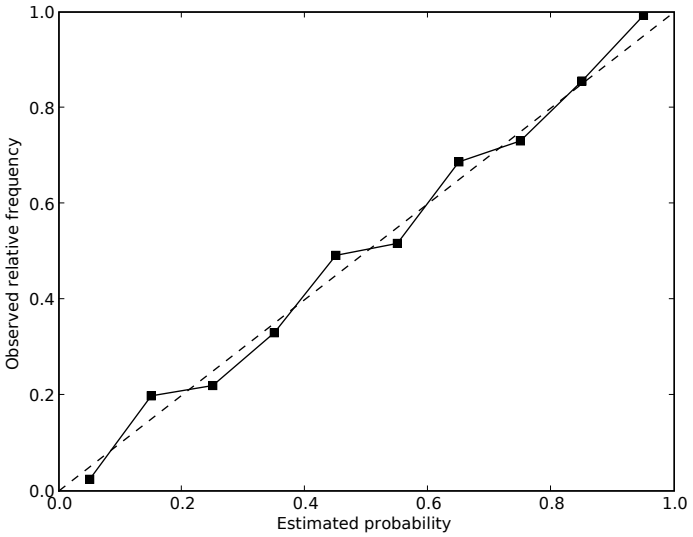


FIG. 8 – Diagramme de fiabilité.

Notations L'ensemble est noté $\mathcal{E}$; il est composé des modèles M^m pour $m \in [\![1, N]\!]$ où $N = |\mathcal{E}|$ est le cardinal de $\mathcal{E}$, donc le nombre de modèles. La concentrations simulée par le modèle M^m à l'instant t et à la position x est notée $M^m_{t,x}$. Une observation à l'instant t et à la position x est notée $O_{t,x}$.

Avec ces notations, l'erreur quadratique moyenne, notée RMSE (« root mean square error ») vaut

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{t,x} \left(M^m_{t,x} - O_{t,x} \right)^2} \tag{26}$$

pour le modèle M^m, où n est le nombre d'observations (soit le nombre de couples (t, x) dans la somme).

La moyenne est notée par une barre verticale suivie de « t », « x » ou « t, x » selon que la moyenne est temporelle, spatiale ou spatio-temporelle (respectivement). Par exemple, la moyenne spatio-temporelle des concentrations du modèle M^m est $\overline{M^m}^{t,x}$.

5.1 Combinaison de modèles

5.1.1 Moyenne d'ensemble

La combinaison la plus simple est la moyenne d'ensemble, notée EM (« ensemble mean ») :

$$\mathrm{EM}_{t,x} = \frac{1}{N} \sum_{m=1}^{N} M^m_{t,x} \, . \tag{27}$$

Une telle combinaison tend vers l'espérance de la concentration en (t, x), lorsque N tend vers l'infini. Si on suppose que le vecteur aléatoire des concentrations a pour espérance les concentrations réelles, la moyenne d'ensemble peut être performante pour la prévision.

Il faut un nombre suffisant de modèles pour que la moyenne d'ensemble converge vers l'espérance.

Si le nombre de modèles est réduit, une solution est de choisir les modèles dans le but d'estimer correctement l'espérance. Ceci fait écho aux méthodes d'échantillonnage des méthodes Monte Carlo (section 2.2). En pratique, un ensemble multi-modèles étant construit sur la base de changements affectant la formulation même du modèle, construire les modèles d'un ensemble peut être délicat d'un point de vue informatique.

Une autre « combinaison » simple est l'ensemble médian EMD :

$$\mathrm{EMD}_{t,x} = \mathrm{median}(M^1_{t,x}, M^2_{t,x}, M^3_{t,x}, \dots M^N_{t,x}) \ . \tag{28}$$

L'ensemble médian présente l'intérêt de filtrer un modèle (ou un petit nombre de modèles) produisant des résultats très différents de la plupart des autres modèles et donc supposés irréalistes.

5.1.2 Superensembles

L'objectif est souvent de produire une prévision de faible erreur quadratique moyenne. L'idée des superensembles [Krishnamurti *et al.*, 2000] est de produire une combinaison linéaire dont les coefficients minimisent l'erreur quadratique sur une période d'apprentissage.

Considérons la combinaison linéaire

$$\mathrm{ELS}_{t,x} = \sum_m \alpha^m M^m_{t,x} \tag{29}$$

où un poids α^m est affecté à chaque modèle. Les poids α^m sont optimaux (pour la cible RMSE) sur toute la période de simulation s'ils minimisent l'erreur quadratique

$$\sum_{t,x} \left[O_{t,x} - \sum_m \alpha^m M^m_{t,x} \right]^2 \ .$$

Les poids n'ont aucune contrainte : ils ne sont pas bornés, ils peuvent être négatifs. De ce fait, les poids ne peuvent donner lieu à aucune interprétation sur la qualité des modèles.

Une version débiaisée est

$$\mathrm{EULS}_{t,x} = \overline{O}^{t,x} + \sum_m \alpha^m \left(M^m_{t,x} - \overline{M^m}^{t,x} \right) \tag{30}$$

où les α^m minimisent

$$\sum_{t,x} \left[O_{t,x} - \overline{O}^{t,x} - \sum_m \alpha^m \left(M^m_{t,x} - \overline{M^m}^{t,x} \right) \right]^2 \ . \tag{31}$$

En prévision, seules les observations du passé sont disponibles. Les poids sont donc calculés sur une période d'apprentissage précédant l'échéance de prévision. La combinaison devient

$$\mathrm{ELS}^{30}_{T,x} = \sum_m \beta^m_T M^m_{T,x} \tag{32}$$

où les poids β_T^m dépendent du temps et minimisent

$$\sum_{t=T-30}^{t=T-1} \sum_x \left[O_{t,x} - \sum_m \beta_T^m M_{t,x}^m \right]^2 ,$$

où $T-1$ est l'instant de l'échéance précédente et $T-30$ est 30e échéance passée. Dans ce cas, la période d'apprentissage $[T-30, T-1]$ est glissante.

5.2 Sélection de modèle et apprentissage statistique

Cette section introduit à des méthodes issues de l'apprentissage statistique (en anglais, « machine learning »). Elles peuvent servir à la sélection de modèles dans un ensemble ou à l'agrégation de modèles (combinaisons linéaires). Contrairement aux super-ensembles, ces méthodes bénéficient d'un cadre mathématique rigoureux, avec des performances contrôlées par des bornes théoriques.

On suppose dans cette section que les observations O_t et les concentrations simulées M_t^m ne dépendent que du temps. On note $l(M_t^m, O_t)$ la mesure de la distance entre la concentration simulée par le modèle M^m et l'observation O_t au temps t. En général, on considère l'erreur quadratique $l(M_t^m, O_t) = (M_t^m - O_t)^2$. On appelle l la fonction de perte.

On introduit aussi la perte cumulée, sur les n prévisions, du modèle M^m

$$L_n^m = \sum_{i=1}^n l(M_i^m, O_i) \ . \tag{33}$$

On cherche à construire une prévision EG dont la perte cumulée $L_n = \sum_{i=1}^n l(\mathrm{EG}_i, O_i)$ soit au plus proche de la perte cumulée du meilleur modèle. On cherche donc à minimiser le regret (externe)

$$R_n = L_n - \min_m L_n^m = \sum_{i=1}^n l(\mathrm{EG}_i, O_i) - \min_m \sum_{i=1}^n l(M_i^m, O_i) \ , \tag{34}$$

sans hypothèse stochastique sur les observations. Tout comme précédemment, EG_t est déterminé sur la base des observations passées $(O_i)_{i<t}$ et des prévisions des modèles $(M_i^m)_{m\in[1,N], i\leq t}$ (aussi appelées avis d'experts).

Un algorithme classique, dit par pondération exponentielle [Vovk, 1990; Cesa-Bianchi, 1997], affecte des poids p_t^m aux modèles, uniforme initialement :

$$p_1^m = \frac{1}{N} \tag{35}$$

et ensuite corrigés selon les pertes :

$$p_t^m = \frac{\exp\left(-\eta \sum_{i=1}^{t-1} l(M_i^m, O_i)\right)}{\sum_{j=1}^N \exp\left(-\eta \sum_{i=1}^{t-1} l(M_i^j, O_i)\right)} \tag{36}$$

où η est un taux d'apprentissage à fixer.

La prévision de l'algorithme au temps t est aléatoire et est notée EGR_t. Elle est tirée, aléatoirement à chaque échéance, parmi les prévisions des modèles selon les probabilités $(p_t^1, p_t^2, \ldots, p_t^N)$.

Si on suppose que les pertes sont bornées, soit $0 \leq l \leq B$, l'espérance $\mathrm{E}(R_n)$ du regret R_n de la prévision EGR_t est bornée. On peut montrer que, quelle que soit la suite des observations,

$$\mathrm{E}(R_n) = \sum_{i=1}^{n} \mathrm{E}(l(\mathrm{EGR}_i, O_i)) - \min_m \sum_{i=1}^{n} l(M_i^m, O_i) \leq \frac{\ln N}{\eta} + \frac{\eta n B^2}{8} \,. \tag{37}$$

En particulier, pour $\eta = \frac{1}{B}\sqrt{\frac{8 \ln N}{n}}$,

$$\mathrm{E}(R_n) \leq B\sqrt{\frac{n \ln N}{2}} \,. \tag{38}$$

Ceci reste vrai quelle que soit la suite de données : la méthode est robuste.

Dans le cas où la fonction de perte est quadratique, c'est-à-dire si $l(M_t^m, O_t) = (M_t^m - O_t)^2$, il est possible de s'affranchir des tirages aléatoires. Les probabilités p_t^m sont considérées comme des poids qui permettent de construire la combinaison EG_t :

$$\mathrm{EG}_t = \sum_{m=1}^{N} p_t^m M_t^m \,. \tag{39}$$

Dans ce cas le regret, par rapport à la meilleure combinaison linéaire (des modèles) constante en temps et dont les poids se somment à 1, est de l'ordre de $B^2\sqrt{n \ln N}$.

Ces méthodes ont l'avantage d'être robustes et d'avoir un cadre mathématique rigoureux. En pratique, pour la prévision de pics d'ozone, par exemple, leurs performances sont inférieures à celles des superensembles. Une stratégie intermédiaire consiste à ajouter à l'ensemble des modèles de nouveaux modèles, dont les superensembles, et d'appliquer à ce nouvel ensemble une méthode d'apprentissage statistique. Puisque les méthodes d'apprentissage statistique garantissent des performances proches du meilleur modèle ou de la meilleure combinaison constante de modèles, les performances des superensembles (dans l'ensemble) sont reproduites, voire améliorées.

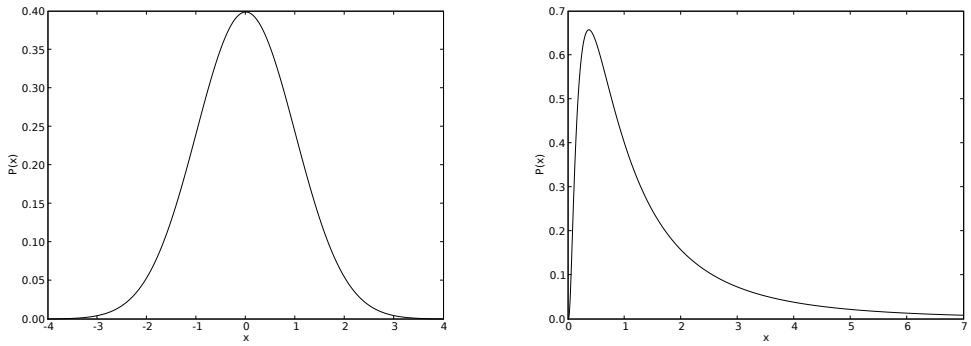
A Lois normale et lognormale

La loi normale $\mathcal{N}(\mu, \sigma^2)$, de moyenne μ et de variance σ^2 a pour densité de probabilité

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \,. \tag{40}$$

La médiane de la distribution est μ.

Lors de tirages aléatoires, les valeurs sont dans l'intervalle $[\mu - \sigma, \mu + \sigma]$ avec une probabilité 0.68, et dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$ avec une probabilité 0.95.



(a) Distribution normale (b) Distribution lognormale

FIG. 9 – Densités de probabilité pour les lois $\mathcal{N}(0,1)$ et $\mathcal{LN}(0,1)$.

Une variable aléatoire X suit une loi lognormale $\mathcal{LN}(\mu,\sigma^2)$ si $\ln X$ suit la loi normale $\mathcal{N}(\mu,\sigma^2)$. La densité de probabilité de la loi est

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) . \tag{41}$$

L'espérance de X est

$$\mathrm{E}(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right) , \tag{42}$$

et sa variance vaut

$$\mathrm{var}(X) = \left(\exp\left(\sigma^2\right) - 1\right) \exp\left(2\mu + \sigma^2\right) . \tag{43}$$

Soient x_i le ie percentile de X et w_i le ie percentile de $\ln X$. Alors

$$x_i = \exp(w_i) . \tag{44}$$

En particulier, la médiane de $\ln X$ étant μ, la médiane de X est $\exp(\mu)$. De même, lors de tirages aléatoires, les valeurs sont dans l'intervalle $[\exp(\mu-\sigma), \exp(\mu+\sigma)]$ avec une probabilité 0.68.

Les densités de probabilité sont représentées à la figure 9.

Références

BEEKMANN, M. et DEROGNAT, C. (2003). Monte Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the atmospheric pollution over the Paris area (ESQUIF) campaign. *Journal of Geophysical Research*, 108(D17):8,559.

CESA-BIANCHI, N. (1997). Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59:392–411.

Chang, J., Brost, R., Isaken, I., Madronich, S., Middleton, P., Stockwell, W. et Walcek, C. (1987). A three-dimensional Eulerian acid deposition model : physical concepts and formulation. *Journal of Geophysical Research*, 92(D12):14,681–14,700.

ESQUIF (2001). Étude et simulation de la qualité de l'air en île de France – rapport final.

Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J. C. *et al.* (2004). Ensemble dispersion forecasting – part I : concept, approach and indicators. *Atmospheric Environment*, 38(28):4,607–4,617.

Gery, M. W., Whitten, G. Z., Killus, J. P. et Dodge, M. C. (1989). A photochemical kinetics mechanism for urban and regional scale computer modeling. *Journal of Geophysical Research*, 94:12,925–12,956.

Gross, A. et Stockwell, W. R. (2003). Comparison of the EMEP, RADM2 and RACM mechanisms. *Journal of Atmospheric Chemistry*, 44:151–170.

Hanna, S. R., Chang, J. C. et Fernau, M. E. (1998). Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmospheric Environment*, 32(21):3,619–3,628.

Hanna, S. R. et Davis, J. M. (2002). Evaluation of a photochemical grid model using estimates of concentration probability density functions. *Atmospheric Environment*, 36:1,793–1,798.

Hanna, S. R., Lu, Z., Frey, H. C., Wheeler, N., Vukovich, J., Arunachalam, S., Fernau, M. et Hansen, D. A. (2001). Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmospheric Environment*, 35(5):891–903.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., T. LaRow, D. B. et Williford, E. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13:4,196–4,216.

Louis, J.-F. (1979). A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology*, 17:187–202.

Madronich, S. (1987). Photodissociation in the atmosphere : 1. actinic flux and the effects of ground reflections and clouds. *Journal of Geophysical Research*, 92(D8):9,740–9,752.

Mallet, V. (2005). *Estimation de l'incertitude et prévision d'ensemble avec un modèle de chimie-transport – Application à la simulation numérique de la qualité de l'air*. Thèse de doctorat, École nationale des ponts et chaussées.

Simpson, D., Winiwarter, W., Börjesson, G., Cinderby, S., Ferreiro, A., Guenther, A., Hewitt, C. N., Janson, R., Khalil, M. A. K., Owen, S., Pierce, T. E., Puxbaum, H., Shearer, M., Skiba, U., Steinbrecher, R., Tarrasón, L. et Öquist, M. G. (1999). Inventorying emissions from nature in Europe. *Journal of Geophysical Research*, 104(D7):8,113–8,152.

Stockwell, W. R., Kirchner, F., Kuhn, M. et Seefeld, S. (1997). A new mechanism for regional atmospheric chemistry modeling. *Journal of Geophysical Research*, 102(D22):25,847–25,879.

Stockwell, W. R., Middleton, P., Chang, J. S. et Tang, X. (1990). The second generation regional acid deposition model chemical mechanism for regional air quality modeling. *Journal of Geophysical Research*, 95(D10):16,343–16,367.

Troen, I. et Mahrt, L. (1986). A simple model of the atmospheric boundary layer ; sensitivity to surface evaporation. *Boundary-Layer Meteorology*, 37:129–148.

Vovk, V. (1990). Aggregating strategies. *In* Kaufmann, M., éditeur : *Proceedings of the 3rd annual workshop on computational learning theory*, pages 371–383, San Mateo, CA.

Wesely, M. L. (1989). Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmospheric Environment*, 23:1,293–1,304.

Wesely, M. L. et Hicks, B. B. (2000). A review of the current status of knowledge on dry deposition. *Atmospheric Environment*, 34:2,261–2,282.

Zhang, L., Brook, J. R. et Vet, R. (2003). A revised parameterization for gaseous dry deposition in air-quality models. *Atmospheric Chemistry and Physics*, 3:2,067–2,082.