Novembre 2010

PERPOL

PERFORMANCES DES OUTILS DE MODELISATION DE LA PLATE FORME VIGIPOL

NOTE MÉTHODOLOGIQUE

En collaboration avec
NUMTECH

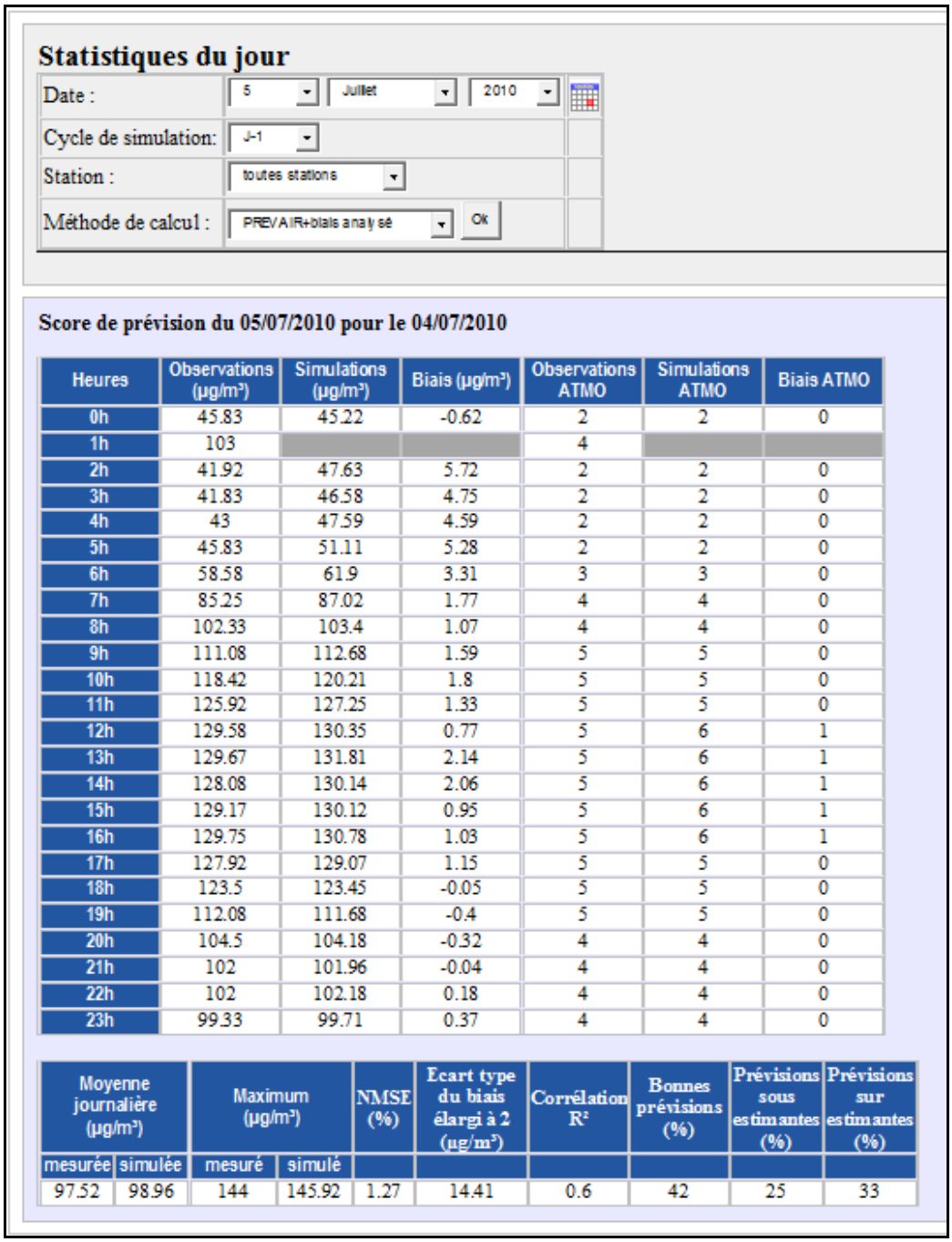
AIRFOBEP
Association pour la Surveillance de la Qualité de l'Air de la Région de l'Etang de Berre et de l'Ouest des Bouches-du-Rhône
Route de la Vierge - 13 500 Martigues - Tel. 04 42 13 01 20 - Fax. 04 42 13 01 29
Site internet: www.airfobep.org - e-mail : airfobep@airfobep.org
Serveur vocal 04 42 49 35 35 (selon tarification téléphonique en vigueur)

1 Introduction et objectifs

Les performances de l'ensemble des cinq plateformes de modélisation développées par NUMTECH pour le compte d'AIRFOBEP (O_3, PM, SO_2, NO_2 et IQA) sont actuellement consultables via l'application SuiviStat. Au cours des dernières semaines, divers dysfonctionnements de l'application ont pu être constatés, poussant ainsi AIRFOBEP et NUMTECH à penser à une mise à jour majeure de celle-ci. Au cours de la revue de projet du 29 juin 2010, il a alors été convenu d'harmoniser l'ensemble des calculs de performances pour chacune des plateformes de modélisation. Nous présenterons ainsi dans cette note l'application SuiviStat telle qu'elle devrait être après harmonisation des méthodologies utilisées et prendrons soin de détailler les spécificités de l'application en fonction du polluant étudié (seuils de concentration utilisés, stations à prendre en compte pour les calculs de performances, etc.). Dans la section 2, nous nous attacherons tout d'abord à définir une méthodologie commune à l'ensemble des plateformes pour les statistiques portant sur une journée uniquement. La section 3 sera quant-à-elle consacrée aux calculs de performances réalisés pour une période de plusieurs jours. Enfin, la section 4 concernera le format des fichiers statistiques à prendre en entrée pour les calculs de performances ainsi que le format des statistiques à afficher sur SuiviStat. Les sections 2, 3 et 4 concernant les plateformes O_3, PM, SO_2, et NO_2, le cas particulier de la plateforme IQA sera décrit dans la section 5.

2 Harmonisation de SuiviStat pour des calculs statistiques journaliers

Suite à la revue de projet du 29 juin 2010, AIRFOBEP a fourni un exemple de la visualisation prévue sous SuiviStat des performances d'une plateforme donnée pour un jour particulier (Figure 1).



Statistiques du jour

Date : 5 Juillet 2010
Cycle de simulation: J-1
Station : toutes stations
Méthode de calcul : PREVAIR+biais analysé Ok

Score de prévision du 05/07/2010 pour le 04/07/2010

Heures	Observations (µg/m³)	Simulations (µg/m³)	Biais (µg/m³)	Observations ATMO	Simulations ATMO	Biais ATMO
0h	45.83	45.22	-0.62	2	2	0
1h	103			4		
2h	41.92	47.63	5.72	2	2	0
3h	41.83	46.58	4.75	2	2	0
4h	43	47.59	4.59	2	2	0
5h	45.83	51.11	5.28	2	2	0
6h	58.58	61.9	3.31	3	3	0
7h	85.25	87.02	1.77	4	4	0
8h	102.33	103.4	1.07	4	4	0
9h	111.08	112.68	1.59	5	5	0
10h	118.42	120.21	1.8	5	5	0
11h	125.92	127.25	1.33	5	5	0
12h	129.58	130.35	0.77	5	6	1
13h	129.67	131.81	2.14	5	6	1
14h	128.08	130.14	2.06	5	6	1
15h	129.17	130.12	0.95	5	6	1
16h	129.75	130.78	1.03	5	6	1
17h	127.92	129.07	1.15	5	5	0
18h	123.5	123.45	-0.05	5	5	0
19h	112.08	111.68	-0.4	5	5	0
20h	104.5	104.18	-0.32	4	4	0
21h	102	101.96	-0.04	4	4	0
22h	102	102.18	0.18	4	4	0
23h	99.33	99.71	0.37	4	4	0

Moyenne journalière (µg/m³)		Maximum (µg/m³)		NMSE (%)	Ecart type du biais élargi à 2 (µg/m³)	Corrélation R^2	Bonnes prévisions (%)	Prévisions sous estimantes (%)	Prévisions sur estimantes (%)
mesurée	simulée	mesuré	simulé						
97.52	98.96	144	145.92	1.27	14.41	0.6	42	25	33

Figure 1 : Interface type prévue pour l'affichage de scores de performances journaliers dans le cadre de l'harmonisation de l'application SuiviStat[1].

[1] Il est envisagé d'alléger l'affichage des noms des différents scores statistiques dans l'ensemble des tableaux de résultats accessibles sur le site de SuiviStat. L'idée générale est d'utiliser des acronymes ou abréviations à la place des dénominations complètes de chaque score (par exemple, « corrélation » pourrait devenir « R^2 » ou encore « COR »). Aussi, il sera possible pour l'utilisateur d'accéder à une page décrivant la signification de l'ensemble des scores (une sorte de glossaire reprenant en partie les définitions données dans ce document) par simple clic sur la case contenant le nom du score pour lequel on désire obtenir des informations.

2.1 Configuration des calculs

Le cadre supérieur (« Statistiques du jour ») de la Figure 1 permet de définir différents paramètres :

- Date : **il s'agit du jour d'observation que l'on désire étudier** (les fichiers stat dans lesquels sont stockées les données observées et simulées étant créés par l'application le lendemain du jour étudié, il sera nécessaire de tenir compte de ce décalage dans la lecture des fichiers stat pour les calculs statistiques : voir section 4). Par défaut, la date affichée est donc celle **de la veille** (dernier jour pour lequel les observations sont disponibles).
- Cycle de simulation : il sera ici possible de choisir si l'on désire évaluer les performances de l'application en mode « prévision » (J, J+1 et J+2 en fonction de la plate-forme considérée) ou « analyse » (calculs effectués pour J-1 en tenant compte ou non de données d'observations pour la météorologie et la pollution).
- Station : en fonction de l'application, il est possible de réaliser des calculs de performances en considérant une seule station du réseau AIRFOBEP, toutes les stations à la fois ou encore certains groupes de stations représentatives de zones urbaines, rurales ou industrielles (cas de l'O_3 uniquement).
- Méthode de calcul : chaque application disposant de différentes étapes de calcul (simulation brute, simulation brute + biais, etc.), il est ici possible de choisir l'étape de calcul dont on désire évaluer les performances.

L'ensemble de ces paramètres peut prendre différentes valeurs en fonction du polluant considéré. Celles-ci sont renseignées dans le Tableau 1.

Notons que dans le cadre de la plate-forme de modélisation de l'O_3, les stations ACGM, AIXA, AIXP, LPNT et SAZE ne seront prises en compte que dans le cas de calculs de performances par station. Celles-ci seront exclues des calculs portant sur l'ensemble des stations ou par groupe de stations.

	O_3	PM	NO_2	SO_2
Date	Quelle que soit l'application, consiste simplement à choisir le jour, le mois et l'année d'intérêt			
Cycle de simulation	J-1, J, J+1 et J+2	J-1, J et J+1	J-1, J, J+1 et J+2	J-1, J, J+1 et J+2
Station	- *ACGM, AIXA, AIXP*, BETG, CRAU, FSCB, ISTR, *LPNT*, MNDM, MRMV, RBRT, *SAZE*, SLPV, SMMR, SRMY, SSLP, VTRL - Urbain (stations en rouge ci-dessus) - Rural (stations en vert ci-dessus) - Industriel (stations en bleu ci-dessus) - Toutes stations (ensemble des stations en couleur ci-dessus)	- ARLS, RBRT, PDBL, MEDE, FSCB, MRMV, MRGV, SLPV, MILE, PSLV - Toutes stations	- ARLS, ISTR, MILE, MRGV, RBRT, SLPV - Toutes stations	- ARLS, BETG, BMGS, CHNF, CLRT, FOLV, FSCB, FSMR, ISTR, MCRN, MEDE, MGTS, MILE, MLVR, MNDM, MPNT, MPTI, MRGV, MRMV, MVTR, PDBC, PDBE, PDBL, PSLV, RBRT, SLPV, SSLP, VTRL - Toutes stations
Méthode de calcul	- PREVAIR - PREVAIR + biais - PREVAIR + biais analysé (J-1)	- ADMS - ADMS + fond - ADMS + fond + biais - ADMS + fond + biais analysé (J-1)	- PREVAIR - PREVAIR + dérivée externe - PREVAIR + dérivée externe analysé (J-1)	- ADMS - ADMS + correction trajectoire des panaches et des taux d'émission (J-1) - ADMS + correction trajectoire des panaches et des taux d'émission analysé (J-1)

Tableau 1 : Valeurs prises par les différents paramètres de configuration de l'application SuiviStat pour un calcul de performances journalières. Pour la plate-forme O_3, les couleurs indiquent le type de station : urbain (en rouge), rural (en vert) ou industriel (en bleu)

2.2 Comparaison heure par heure des résultats de simulation

Le premier tableau du cadre (« Score de prévision du 05/07/2010 pour le 04/07/2010 ») de la Figure 1 présente des résultats obtenus heure par heure à la fois pour les observations et les simulations des concentrations de polluants. Les différents paramètres évalués sont :

- Observations (µg/m³) :
  - Si une seule station est sélectionnée, il s'agira de la valeur observée à l'heure donnée dans la première colonne (heure TU). Une case grise indique une donnée invalide.
  - Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), le nombre reporté correspondra à la moyenne des concentrations observées en chacune des stations. **Ici, seules les données d'observations pour lesquelles les simulations correspondantes sont disponibles seront utilisées.**
- Simulations (µg/m³) :
  - Si une seule station est sélectionnée, il s'agira de la valeur simulée à l'heure donnée dans la première colonne (heure TU). Une case grise indique une absence de calcul pour cette heure.
  - Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), le nombre reporté correspondra à la moyenne des concentrations simulées en chacune des stations. **Ici, seules les données de simulations pour lesquelles les observations correspondantes sont disponibles seront utilisées.**
- Biais (µg/m³) : il s'agit de l'erreur moyenne commise sur l'évaluation des concentrations observées pour l'heure étudiée. Le biais est calculé selon la formule suivante :

$$Biais = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)$$

  Où O_i et S_i représentent respectivement une concentration horaire observée dans la classe et la simulation correspondante, et N le nombre total de couples (O_i, S_i) utilisés pour cette classe. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles seront considérées.**
- Observations ATMO : il s'agit de la valeur du sous-indice ATMO associée à la valeur de concentration reportée dans la colonne « Observations (µg/m³) ». Les valeurs des sous-indices ATMO associés à chaque polluant sont renseignées dans les Tableaux 2 à 5.
- Simulations ATMO : il s'agit de la valeur du sous-indice ATMO associée à la valeur de concentration reportée dans la colonne « Simulations (µg/m³) ». Les valeurs des sous-indices ATMO associés à chaque polluant sont renseignées dans les Tableaux 2 à 5.
- Biais ATMO : il s'agit de l'erreur moyenne commise sur l'évaluation des sous-indices ATMO observés pour l'heure étudiée. La formule utilisée est la même que celle donnée ci-dessus. **Notons que par construction, pour le calcul de ce paramètre, seules les journées de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

Les Tableaux 2 à 5 présentent les valeurs du sous-indice ATMO associées aux concentrations horaires pour chacun des polluants. Notons que le sous-indice ATMO est censé être associé à un indice de la concentration en un polluant au cours de la journée (concentration moyenne journalière pour les PM et concentration maximale journalière pour O_3, SO_2 et NO_2) mais qu'ici, on utilise de façon abusive les sous-indices ATMO en les appliquant sur des concentrations horaires.

Sous-indice PM	Seuil min. (inclus) en µg/m³	Seuil max. (exclus) en µg/m³
1	0	10
2	10	20
3	20	30
4	30	40
5	40	50
6	50	65
7	65	80
8	80	100
9	100	125
10	125	

Tableau 2 : Grille de calcul du sous-indice ATMO pour les concentrations horaires en PM.

Sous-indice NO_2	Seuil min. (inclus) en µg/m³	Seuil max. (exclus) en µg/m³
1	0	30
2	30	55
3	55	85
4	85	110
5	110	135
6	135	165
7	165	200
8	200	275
9	275	400
10	400	

Tableau 3 : Grille de calcul du sous-indice ATMO pour les concentrations horaires en NO_2.

Sous-indice O_3	Seuil min. (inclus) en µg/m³	Seuil max. (exclus) en µg/m³
1	0	30
2	30	55
3	55	80
4	80	105
5	105	130
6	130	150
7	150	180
8	180	210
9	210	240
10	240	

Tableau 4 : Grille de calcul du sous-indice ATMO pour les concentrations horaires en O_3.

Sous-indice SO_2	Seuil min. (inclus) en µg/m³	Seuil max. (exclus) en µg/m³
1	0	40
2	40	80
3	80	120
4	120	160
5	160	200
6	200	250
7	250	300
8	300	400
9	400	500
10	500	

Tableau 5 : Grille de calcul du sous-indice ATMO pour les concentrations horaires en SO_2.

2.3 Comparaison sur l'ensemble de la journée

Le deuxième tableau du cadre (« Score de prévision du 05/07/2010 pour le 04/07/2010 ») de la Figure 1 présente des scores de performances portant sur l'intégralité de la journée. Les différents paramètres évalués sont :

- Moyenne journalière mesurée (µg/m³). Si on note O_i une concentration horaire observée, la moyenne journalière M sera :

$$M = \frac{1}{N}\sum_{i=1}^{N} O_i$$

  Où N représente le nombre d'échéances horaires utilisées (i.e. auxquelles des observations étaient disponibles) pour réaliser le calcul de moyenne. **Ici, seules les données d'observations pour lesquelles les simulations correspondantes sont disponibles seront utilisées.**
- Moyenne journalière simulée (µg/m³). Si on note S_i une concentration horaire simulée, la moyenne M sera :

$$M = \frac{1}{N}\sum_{i=1}^{N} S_i$$

  Où N représente le nombre d'échéances horaires utilisées (i.e. auxquelles des simulations étaient disponibles) pour réaliser le calcul de moyenne. **Ici, seules les données de simulations pour lesquelles les observations correspondantes sont disponibles seront utilisées.**
- Maximum mesuré (µg/m³) :
  - Si une seule station est sélectionnée, il s'agira de la valeur horaire maximale observée à cette station.
  - Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), le nombre reporté correspondra à la concentration maximale horaire observée toutes stations confondues.

  Ici, seules les données d'observations pour lesquelles les simulations correspondantes sont disponibles seront utilisées.
- Maximum simulé (µg/m³) :
  - Si une seule station est sélectionnée, il s'agira de la valeur horaire maximale simulée à cette station.
  - Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), le nombre reporté correspondra à la concentration maximale horaire simulée toutes stations confondues.

  Ici, seules les données de simulations pour lesquelles les observations correspondantes sont disponibles seront utilisées.
- NMSE (%) : il s'agit de l'erreur quadratique normalisée. Sa formule est la suivante :

$$NMSE = 100 \times \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N} S_i \times O_i}$$

  Où N représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances horaires de la journée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Ecart-type du biais élargi à 2 (µg/m³) : il s'agit de l'écart-type (multiplié par 2) du biais entre observations et simulations horaires. Sa formule est la suivante :

$$\text{Ecart type du biais élargi à 2} = 2 \times \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left((S_i - O_i) - (\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i))\right)^2}$$

Où *N* représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances horaires de la journée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Corrélation R^2 : il s'agit du coefficient de corrélation de Pearson r (corrélation linéaire) élevé au carré que l'on calcule entre les observations et les simulations horaires. Sa formule est la suivante :

$$\textit{Coefficient de corrélation } r = \frac{\sum_{i=1}^{N}\left(O_i - \frac{1}{N}\sum_{i=1}^{N}O_i\right)\times\left(S_i - \frac{1}{N}\sum_{i=1}^{N}S_i\right)}{\sqrt{\sum_{i=1}^{N}\left(O_i - \frac{1}{N}\sum_{i=1}^{N}O_i\right)^2}\times\sqrt{\sum_{i=1}^{N}\left(S_i - \frac{1}{N}\sum_{i=1}^{N}S_i\right)^2}}$$

$$\textit{Corrélation } R^2 = r^2$$

Où *N* représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances horaires de la journée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Bonnes prévisions (%), prévisions sous-estimantes (%) et prévisions sur-estimantes (%) du sous-indice ATMO : ces paramètres sont relatifs aux prévisions du sous-indice ATMO associé au polluant considéré. Le pourcentage de bonnes prévisions est défini comme le nombre d'échéances horaires pour lesquelles le biais ATMO est égal à 0, rapporté au nombre d'échéances horaires de la journée pour lesquelles simulations ATMO et observations ATMO sont disponibles.

$$\textit{Bonnes prévisions (\%)} = 100 \times \frac{1}{N}\sum_{i=1}^{N} previ_i \quad \text{avec } previ_i = 1 \text{ si } \textit{Biais ATMO}_i = 0$$

$$previ_i = 0 \text{ sinon.}$$

Où *N* représente le nombre d'échéances horaires utilisées (i.e auxquelles le biais ATMO était disponible), et $previ_i$ est un indicateur permettant de comptabiliser le nombre de biais ATMO égaux à 0.

De la même façon, le pourcentage de prévisions sous-estimantes (respectivement sur-estimantes) est défini comme le nombre d'échéances horaires pour lesquelles le biais ATMO est négatif (respectivement positif) rapporté au nombre d'échéances horaires de la journée pour lesquelles simulations ATMO et observations ATMO sont disponibles. **Notons que par construction, pour le calcul de ces paramètres, seules les échéances horaires de la journée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

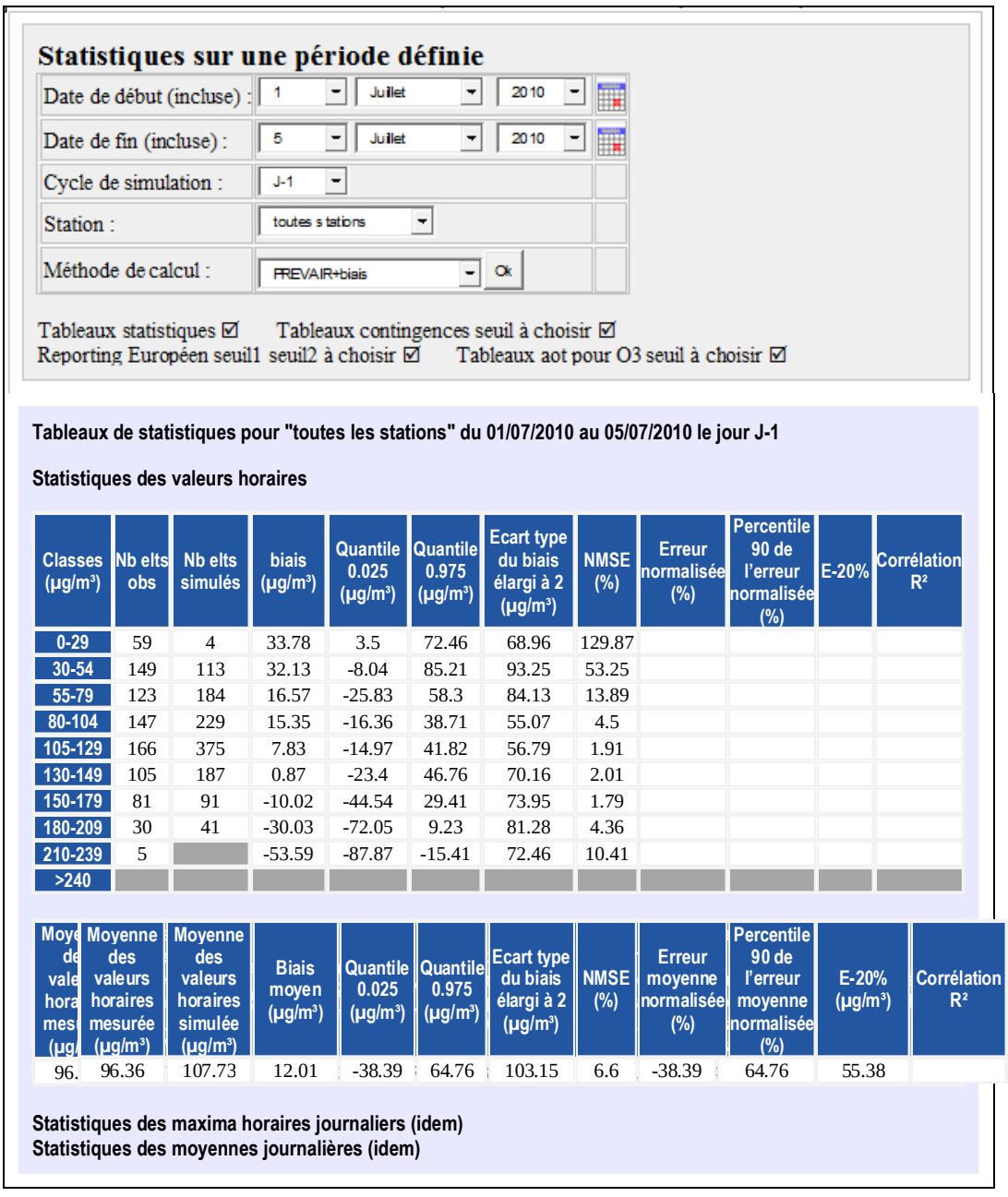
ATTENTION, pour l'ensemble de ces scores :

- ***Si une seule station est sélectionnée, le calcul est réalisé sur l'ensemble des heures disponibles à la station pour la journée : N a pour valeur maximale 24 (si l'ensemble des observations et simulations de 0h à 23h sont disponibles à la station). Les données prises en compte sont celles figurant dans le Tableau 1.***

- ***Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), le calcul est réalisé sur l'ensemble des heures disponibles pour toutes les stations pour la journée : N a pour valeur maximale 24 x $N_{stations}$, avec $N_{stations}$ le nombre de stations (si l'ensemble des observations et simulations de 0h à 23h sont disponibles en chaque station). Les données prises en compte ne sont donc pas celles figurant dans le Tableau 1, qui sont des valeurs moyennées sur les stations. Ceci signifie qu'on ne calcule pas des scores sur des valeurs horaires au préalable moyennées sur les stations, mais bien que l'on calcule des scores sur toutes les échéances et les stations d'un coup.***

3 Harmonisation de SuiviStat pour des calculs statistiques sur une période

Suite à la revue de projet du 29 juin 2010, AIRFOBEP a fourni un exemple de la visualisation prévue sous SuiviStat des performances d'une plateforme donnée pour une période définie (Figure 2).



Statistiques sur une période définie

Date de début (incluse) :	1	Juillet	2010
Date de fin (incluse) :	5	Juillet	2010
Cycle de simulation :	J-1		
Station :	toutes stations		
Méthode de calcul :	PREVAIR+biais	Ok	

Tableaux statistiques ☑ Tableaux contingences seuil à choisir ☑
Reporting Européen seuil1 seuil2 à choisir ☑ Tableaux aot pour O3 seuil à choisir ☑

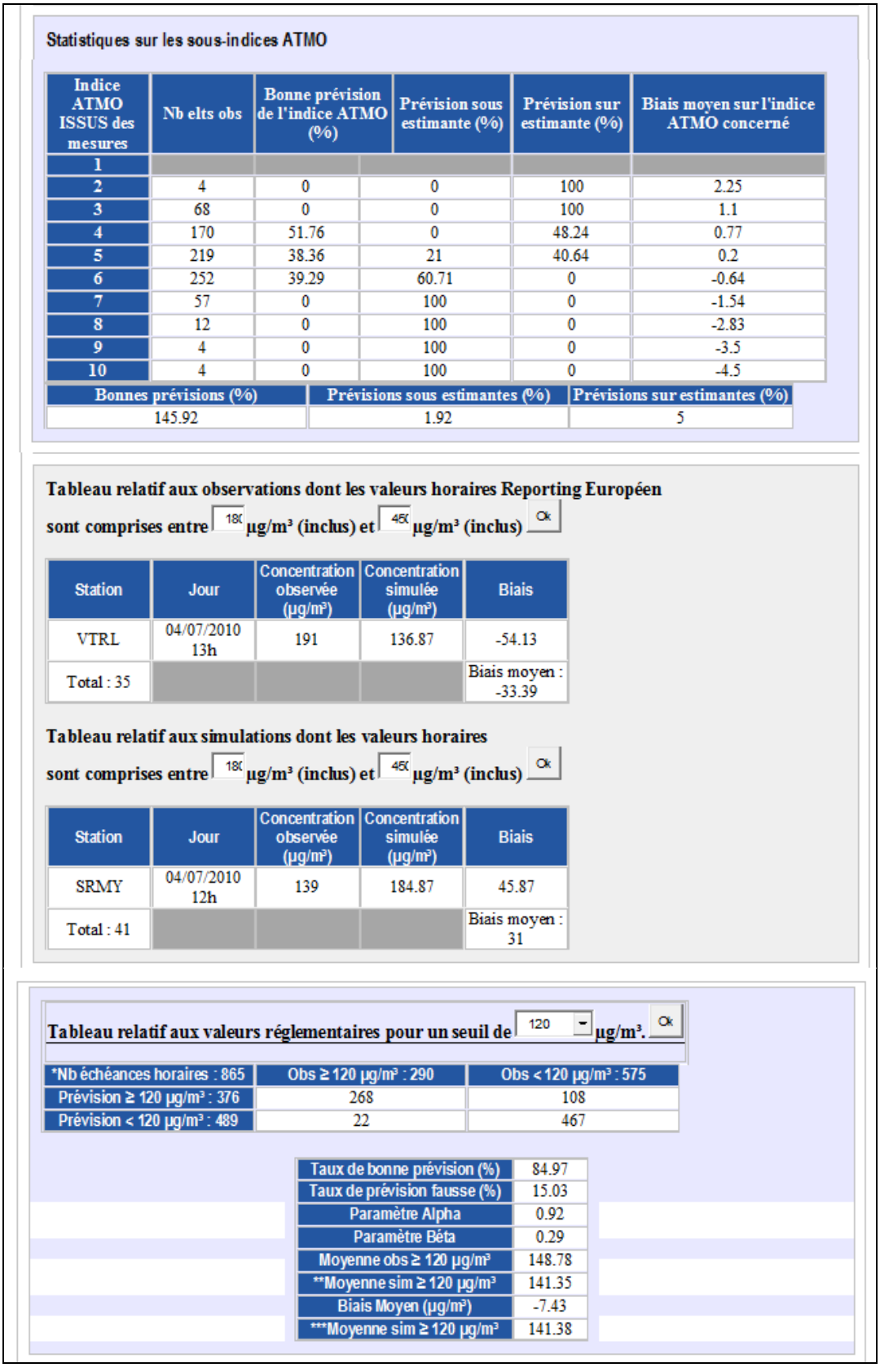
Tableaux de statistiques pour "toutes les stations" du 01/07/2010 au 05/07/2010 le jour J-1

Statistiques des valeurs horaires

Classes (µg/m³)	Nb elts obs	Nb elts simulés	biais (µg/m³)	Quantile 0.025 (µg/m³)	Quantile 0.975 (µg/m³)	Ecart type du biais élargi à 2 (µg/m³)	NMSE (%)	Erreur normalisée (%)	Percentile 90 de l'erreur normalisée (%)	E-20%	Corrélation R²
0-29	59	4	33.78	3.5	72.46	68.96	129.87				
30-54	149	113	32.13	-8.04	85.21	93.25	53.25				
55-79	123	184	16.57	-25.83	58.3	84.13	13.89				
80-104	147	229	15.35	-16.36	38.71	55.07	4.5				
105-129	166	375	7.83	-14.97	41.82	56.79	1.91				
130-149	105	187	0.87	-23.4	46.76	70.16	2.01				
150-179	81	91	-10.02	-44.54	29.41	73.95	1.79				
180-209	30	41	-30.03	-72.05	9.23	81.28	4.36				
210-239	5		-53.59	-87.87	-15.41	72.46	10.41				
>240											

Moyenne de valeurs horaires mesurées (µg/m³)	Moyenne des valeurs horaires mesurée (µg/m³)	Moyenne des valeurs horaires simulée (µg/m³)	Biais moyen (µg/m³)	Quantile 0.025 (µg/m³)	Quantile 0.975 (µg/m³)	Ecart type du biais élargi à 2 (µg/m³)	NMSE (%)	Erreur moyenne normalisée (%)	Percentile 90 de l'erreur moyenne normalisée (%)	E-20% (µg/m³)	Corrélation R²
96.	96.36	107.73	12.01	-38.39	64.76	103.15	6.6	-38.39	64.76	55.38	

Statistiques des maxima horaires journaliers (idem)
Statistiques des moyennes journalières (idem)



Statistiques sur les sous-indices ATMO

Indice ATMO ISSUS des mesures	Nb elts obs	Bonne prévision de l'indice ATMO (%)	Prévision sous estimante (%)	Prévision sur estimante (%)	Biais moyen sur l'indice ATMO concerné
1					
2	4	0	0	100	2.25
3	68	0	0	100	1.1
4	170	51.76	0	48.24	0.77
5	219	38.36	21	40.64	0.2
6	252	39.29	60.71	0	-0.64
7	57	0	100	0	-1.54
8	12	0	100	0	-2.83
9	4	0	100	0	-3.5
10	4	0	100	0	-4.5

Bonnes prévisions (%)	Prévisions sous estimantes (%)	Prévisions sur estimantes (%)
145.92	1.92	5

Tableau relatif aux observations dont les valeurs horaires Reporting Européen sont comprises entre 180 µg/m³ (inclus) et 450 µg/m³ (inclus) Ok

Station	Jour	Concentration observée (µg/m³)	Concentration simulée (µg/m³)	Biais
VTRL	04/07/2010 13h	191	136.87	-54.13
Total : 35				Biais moyen : -33.39

Tableau relatif aux simulations dont les valeurs horaires sont comprises entre 180 µg/m³ (inclus) et 450 µg/m³ (inclus) Ok

Station	Jour	Concentration observée (µg/m³)	Concentration simulée (µg/m³)	Biais
SRMY	04/07/2010 12h	139	184.87	45.87
Total : 41				Biais moyen : 31

Tableau relatif aux valeurs réglementaires pour un seuil de 120 µg/m³. Ok

*Nb échéances horaires : 865	Obs ≥ 120 µg/m³ : 290	Obs < 120 µg/m³ : 575
Prévision ≥ 120 µg/m³ : 376	268	108
Prévision < 120 µg/m³ : 489	22	467

Taux de bonne prévision (%)	84.97
Taux de prévision fausse (%)	15.03
Paramètre Alpha	0.92
Paramètre Béta	0.29
Moyenne obs ≥ 120 µg/m³	148.78
**Moyenne sim ≥ 120 µg/m³	141.35
Biais Moyen (µg/m³)	-7.43
***Moyenne sim ≥ 120 µg/m³	141.38

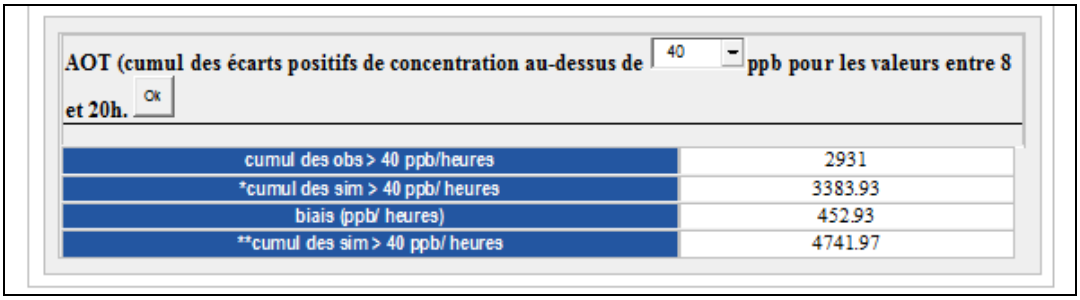


Figure 2 : Interface type prévue pour l'affichage de scores de performances sur une période dans le cadre de l'harmonisation de l'application SuiviStat.

3.1 Configuration des calculs

Le cadre supérieur (« Statistiques sur une période définie ») de la Figure 2 permet de définir différents paramètres :

- Date de début (incluse) : il s'agit du premier jour **d'observation** (inclus) de la période que l'on désire étudier.
- Date de fin (incluse) : il s'agit du dernier jour **d'observation** (inclus) de la période que l'on désire étudier (les fichiers stat dans lesquels sont stockées les données observées et simulées étant créés par l'application le lendemain du jour étudié, il sera nécessaire de tenir compte de ce décalage dans la lecture des fichiers stat pour les calculs statistiques : voir section 4)
- Cycle de simulation (voir Tableau 1) : il sera ici possible de choisir si l'on désire évaluer les performances de l'application en mode « prévision » (J, J+1 et J+2 en fonction de la plate-forme considérée) ou « analyse » (calculs effectués pour J-1 en tenant compte ou non de données d'observations pour la météorologie et la pollution).
- Station (voir Tableau 1) : en fonction de l'application, il est possible de réaliser des calculs de performances en considérant une seule station du réseau AIRFOBEP, toutes les stations à la fois ou encore certains groupes de stations représentatives de zones urbaines, rurales ou industrielles (cas de l'O_3 uniquement).
- Méthode de calcul (voir Tableau 1) : chaque application disposant de différentes étapes de calcul (simulation brute, simulation brute + biais, etc.), il est ici possible de choisir l'étape de calcul dont on désire évaluer les performances.
- Les cases « Tableaux statistiques », « Tableaux contingence seuil à choisir », « Reporting européen seuil1 seuil2 à choisir » et « Tableaux AOT pour O_3 seuil à choisir » (uniquement pour la plateforme O_3) correspondent aux types de calculs de performances que l'on désire effectuer et afficher.

3.2 Case « Tableaux statistiques »

Lorsque la case « Tableaux statistiques » est cochée, les résultats présentés sur la Figure 2 (cadres « Statistiques des valeurs horaires », « Statistiques des maxima horaires journaliers », « Statistiques des moyennes journalières » et « Statistiques sur les sous-indices ATMO ») sont obtenus.

3.2.1 Calculs statistiques sur les valeurs de concentrations horaires

Tableau contenant des calculs de performances par classe de concentration horaire

Le premier tableau concerne la comparaison des données observées et simulées en fonction de la classe de concentration de la concentration horaire observée. Les classes de concentrations sont prédéfinies (colonne « Classes (µg/m³) ») pour chacun des polluants étudiés, ces classes étant en fait celles qui ont été définies pour calculer les valeurs de sous-indice ATMO de chaque polluant et renseignées dans les Tableaux 2 à 5.

Ce tableau contient, pour chaque classe de concentration horaire observée :

- Nb elts obs : il s'agit du nombre de concentrations horaires observées tombant dans la classe. **Ici, seules les données d'observations pour lesquelles les simulations correspondantes sont disponibles seront utilisées pour remplir cette colonne du tableau.**
- Nb elts simulés : il s'agit du nombre de concentrations horaires simulées tombant dans la classe. Ce classement est indépendant de celui des observations et figure uniquement à titre indicatif. **Ici, seules les données de simulations pour lesquelles les observations correspondantes sont disponibles seront utilisées pour remplir cette colonne du tableau.**
- Biais (µg/m³) : il s'agit de l'erreur moyenne commise sur l'évaluation des concentrations horaires observées pour la classe de concentrations. Le biais est calculé selon la formule suivante :

$$Biais = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)$$

  Où O_i et S_i représentent respectivement une concentration horaire observée dans la classe et la simulation correspondante, et N le nombre total de couples (O_i, S_i) utilisés pour cette classe. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles seront considérées.**
- Quantile 0.025 (µg/m³) et Quantile 0.975 (µg/m³) : ces paramètres portent sur le biais existant entre concentration horaire simulée et observée pour la classe de concentrations étudiée. Le quantile 0.025 indique la valeur de biais telle que 2.5% des autres biais lui sont inférieurs, tandis que le quantile 0.975 indique la valeur de biais telle que 2.5% des autres biais lui sont supérieurs. Pour déterminer ces quantiles, il est nécessaire de classer par ordre croissant toutes les valeurs de biais dont on dispose pour la classe et de déterminer le rang de chaque valeur. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles seront considérées.**
- Erreur normalisée : il s'agit de l'erreur moyenne (en %) commise sur l'évaluation des concentrations horaires normalisée par la moyenne des concentrations observées. Elle est calculée comme suit :

$$Erreur\ normalisée = 100 \times \frac{1}{N} \sum_{i=1}^{N} \frac{S_i - O_i}{O_i}$$

Où O_i et S_i représentent respectivement une concentration horaire observée dans la classe et la simulation correspondante, et N le nombre total de couples (O_i, S_i) utilisés pour cette classe. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles seront considérées.**

- Percentile 90 de l'erreur normalisée (%) : il s'agit de la valeur (en %) de l'erreur normalisée en dessous de laquelle se situent 90% des valeurs d'erreurs normalisées.
- E-20% : il s'agit du pourcentage d'heures pour lesquelles les concentrations horaires simulées et observées diffèrent de moins de 20% de la valeur de concentration horaire observée.

$$E\text{-}20\% = 100 \times \frac{1}{N} \sum_{i=1}^{N} erreur_i \quad \text{avec } erreur_i = 1 \text{ si } \frac{|O_i - S_i|}{O_i} \leq 0.2$$

$$erreur_i = 0 \text{ sinon.}$$

Où O_i et S_i représentent respectivement une concentration horaire observée dans la classe et la simulation correspondante, N le nombre total de couples (O_i, S_i) utilisés pour cette classe et $erreur_i$ est un indicateur permettant de comptabiliser le nombre d'erreurs en valeur absolue inférieures à 20% de la concentration horaire observée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Les paramètres "Ecart-type du biais élargi à 2 (µg/m³)", "NMSE (%)" et "Corrélation R²", sont ceux définis dans la section 2.3. **Notons que par construction, pour le calcul de ces paramètres, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

Tableau contenant des calculs de performances sur la concentration horaire

Le second tableau contient les paramètres suivants :

- Moyenne des valeurs horaires mesurées (µg/m³) : il s'agit de la moyenne des valeurs horaires de concentration observées sur la période. Si on note O_i une concentration horaire observée, la moyenne journalière M sera :

$$M = \frac{1}{N} \sum_{i=1}^{N} O_i$$

Où N représente le nombre d'échéances horaires utilisées (i.e. auxquelles des observations étaient disponibles) pour réaliser le calcul de moyenne. **Ici, seules les données d'observations pour lesquelles les simulations correspondantes sont disponibles seront utilisées.**

- Moyenne des valeurs horaires simulées (µg/m³) : il s'agit de la moyenne des valeurs horaires de concentration simulées sur la période. Si on note S_i une concentration horaire simulée, la moyenne journalière M sera :

$$M = \frac{1}{N}\sum_{i=1}^{N} S_i$$

Où *N* représente le nombre d'échéances horaires utilisées (i.e. auxquelles des simulations étaient disponibles) pour réaliser le calcul de moyenne. **Ici, seules les données de simulations pour lesquelles les observations correspondantes sont disponibles seront utilisées.**

- Biais moyen (µg/m³) : il s'agit de l'erreur moyenne commise sur l'évaluation des concentrations horaires. Le biais moyen est calculé selon la formule :

$$\textit{Biais moyen} = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)$$

Où *N* représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Quantile 0.025 (µg/m³) et Quantile 0.975 (µg/m³) : ces paramètres portent sur le biais existant entre concentration horaire simulée et observée. Le quantile 0.025 indique la valeur de biais telle que 2.5% des autres biais lui sont inférieurs, tandis que le quantile 0.975 indique la valeur de biais telle que 2.5% des autres biais lui sont supérieurs. Pour déterminer ces quantiles, il est nécessaire de classer par ordre croissant toutes les valeurs de biais dont on dispose et de déterminer le rang de chaque valeur. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**
- Ecart-type du biais élargi à 2 (µg/m³) : il s'agit de l'écart-type (multiplié par 2) du biais entre simulations et observations horaires. Sa formule est la suivante :

$$\text{Ecart type du biais élargi à 2} = 2\times\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((S_i - O_i) - (\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i))\right)^2}$$

Où *N* représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- NMSE (%) : il s'agit de l'erreur quadratique normalisée, de formule :

$$NMSE = 100\times\frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N} S_i \times O_i}$$

Où *N* représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Erreur moyenne normalisée : il s'agit de l'erreur moyenne (en %) commise sur l'évaluation des concentrations horaires normalisée par la moyenne des concentrations observées. Elle est calculée comme suit :

$$Erreur\ moyenne\ normalisée = 100 \times \frac{1}{N} \sum_{i=1}^{N} \frac{S_i - O_i}{O_i}$$

Où O_i et S_i représentent respectivement une concentration horaire observée dans la classe et la simulation correspondante, et N le nombre total de couples (O_i, S_i) utilisés pour cette classe. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles seront considérées.**

- Percentile 90 de l'erreur normalisée (%) : il s'agit de la valeur (en %) de l'erreur normalisée en dessous de laquelle se situent 90% des valeurs d'erreurs normalisées.
- E-20% : il s'agit du pourcentage d'heures pour lesquelles les concentrations simulées et observées diffèrent de moins de 20% de la valeur de concentration horaire observée.

$$E\text{-}20\% = 100 \times \frac{1}{N} \sum_{i=1}^{N} erreur_i \quad \text{avec } erreur_i = 1 \text{ si } \frac{|O_i - S_i|}{O_i} \leq 0.2$$

$$erreur_i = 0 \text{ sinon.}$$

Où O_i et S_i représentent respectivement une concentration horaire observée et simulée, N le nombre total de couples (O_i, S_i) utilisés et $erreur_i$ est un indicateur permettant de comptabiliser le nombre d'erreurs en valeur absolue inférieures à 20% de la concentration horaire observée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

- Corrélation R^2 : il s'agit du coefficient de corrélation de Pearson r (corrélation linéaire) élevé au carré que l'on calcule entre les observations et les simulations horaires. Sa formule est la suivante :

$$Coefficient\ de\ corrélation\ r = \frac{\sum_{i=1}^{N} \left(O_i - \frac{1}{N} \sum_{i=1}^{N} O_i \right) \times \left(S_i - \frac{1}{N} \sum_{i=1}^{N} S_i \right)}{\sqrt{\sum_{i=1}^{N} \left(O_i - \frac{1}{N} \sum_{i=1}^{N} O_i \right)^2} \times \sqrt{\sum_{i=1}^{N} \left(S_i - \frac{1}{N} \sum_{i=1}^{N} S_i \right)^2}}$$

$$Corrélation\ R^2 = r^2$$

Où N représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une concentration horaire observée et simulée. **Notons que par construction, pour le calcul de ce paramètre, seules les échéances de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

ATTENTION, pour l'ensemble de ces scores :

- ***Si une seule station est sélectionnée, le calcul est réalisé sur l'ensemble des heures disponibles à la station sur la période : N a pour valeur maximale 24 x N_{jours}, avec N_{jours} le nombre de jours considérés.***
- ***Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), le calcul est réalisé sur l'ensemble des heures disponibles pour toutes les stations sur la période : N a pour valeur maximale 24 x***

$N_{stations}$ x N_{jours}, avec $N_{stations}$ le nombre de stations et N_{jours} le nombre de jours considérés.

3.2.2 Calculs statistiques sur la concentration moyenne journalière

Cette partie est identique à la précédente mais porte sur la concentration moyenne journalière. Il est donc nécessaire de calculer les concentrations moyennes journalières observées et simulées.

- Si une seule station est sélectionnée, on calcule la concentration moyenne à la station pour chaque journée de la période.
- Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel) pour la plate-forme O_3, on calcule une concentration moyenne pour chaque station et chaque journée de la période.

ATTENTION, pour l'ensemble des scores :

- ***Si une seule station est sélectionnée, le calcul est réalisé sur l'ensemble des journées disponibles à la station sur la période : N a pour valeur maximale N_{jours}, avec N_{jours} le nombre de jours considérés.***
- ***Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel) pour la plate-forme O_3, le calcul est réalisé sur l'ensemble des journées disponibles pour toutes les stations sur la période : N a pour valeur maximale $N_{stations}$ x N_{jours}, avec $N_{stations}$ le nombre de stations et N_{jours} le nombre de jours considérés.***

3.2.3 Calculs statistiques sur le maximum horaire journalier

De même, cette partie est identique à celle concernant les statistiques sur la concentration horaire mais porte sur la concentration maximale journalière. Il est donc nécessaire de calculer les concentrations maximales journalières observées et simulées.

- Si une seule station est sélectionnée, on calcule la concentration maximale à la station pour chaque journée de la période.
- Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), on calcule une concentration maximale pour chaque station et chaque journée de la période.

ATTENTION, pour l'ensemble des scores :

- ***Si une seule station est sélectionnée, le calcul est réalisé sur l'ensemble des journées disponibles à la station sur la période : N a pour valeur maximale N_{jours}, avec N_{jours} le nombre de jours considérés.***
- ***Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel) pour la plate-forme O_3, le calcul est réalisé sur l'ensemble des journées disponibles pour toutes les stations sur la période : N a pour valeur maximale $N_{stations}$ x N_{jours}, avec $N_{stations}$ le nombre de stations et N_{jours} le nombre de jours considérés.***

3.2.4 Calculs statistiques sur le sous-indice ATMO

Le cadre suivant de la Figure 2 contient des scores de performances portant sur la prévision des sous-indices ATMO associés au polluant étudié. Les sous-indices ATMO sont calculés à partir d'un indice de la concentration du polluant au cours de la journée, cet indice pouvant varier d'un polluant à un autre. Dans le cas des PM, le sous-indice ATMO porte sur la concentration moyenne journalière. Pour les polluants O_3, NO_2 et SO_2, la valeur de sous-indice ATMO porte sur la concentration maximale journalière. Il est donc nécessaire d'associer un sous-indice ATMO observé et simulé à chaque journée. Les valeurs des sous-indices ATMO associés à chaque polluant sont renseignées dans les Tableaux 2 à 5.

- Si une seule station est sélectionnée, on associe un sous-indice ATMO à chaque journée de la période**.**
- Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel pour la plate-forme O_3), on associe un sous-indice ATMO à chaque journée de la période pour chaque station.

Les scores sont tout d'abord répartis par valeur du sous-indice ATMO observé (valeurs définies dans les Tableaux 2 à 5 en fonction du polluant). Pour chaque valeur de sous-indice ATMO observé, les paramètres renseignés sont les suivants :

- Nb elts obs : il s'agit du nombre de sous-indices ATMO observés tombant dans la classe considérée. **Notons que pour le calcul de ce paramètres, seules les journées de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**
- Bonne prévision de l'indice ATMO (%), prévision sous-estimante (%) et prévision sur-estimante. (%) Ces paramètres sont calculés tels que définis dans la section 2.3. **Notons que par construction, pour le calcul de ces paramètres, seules les journées de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**
- Biais moyen sur l'indice ATMO concerné. Ce score indique l'erreur moyenne commise pour les sous-indices ATMO observés. **Notons que par construction, pour le calcul de ce paramètre, seules les journées de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

Une ligne supplémentaire du tableau indique différents paramètres : Bonnes prévisions (%), prévisions sous-estimantes (%) et prévisions sur-estimantes (%). Il s'agit des mêmes calculs que précédemment mais sans distinction de classes des sous-indices ATMO.

ATTENTION, pour l'ensemble des scores :

- ***Si une seule station est sélectionnée, le calcul est réalisé sur l'ensemble des journées disponibles à la station sur la période : N a pour valeur maximale N_{jours}, avec N_{jours} le nombre de jours considérés.***
- ***Si le calcul porte sur toutes les stations ou sur un groupe particulier (urbain, rural ou industriel) pour la plate-forme O_3, le calcul est réalisé sur l'ensemble des journées disponibles pour toutes les stations sur la période : N a pour valeur maximale $N_{stations}$ x N_{jours}, avec $N_{stations}$ le nombre de stations et N_{jours} le nombre de jours considérés.***

3.3 Case « Reporting Européen seuil1 seuil2 à choisir »

Le cadre suivant permet un affichage de la liste des valeurs horaires (d'abord les observations puis les simulations) qui, au cours de la période sélectionnée, ont été comprises entre deux seuils de concentration choisis. La possibilité est laissée à l'utilisateur de rentrer ces deux seuils à la main (exprimés en µg/m³). Par défaut, les seuils minimums et maximums fixés pour les différents polluants figurent dans le Tableau 6.

Polluant	Seuil min. (inclus : ≥) en µg/m³	Seuil max. (inclus : ≤) en µg/m³
PM	80	Maximum accepté par BADOS
O_3	180	Maximum accepté par BADOS
NO_2	200	Maximum accepté par BADOS
SO_2	300	Maximum accepté par BADOS

Tableau 6 : Seuils de concentration pris par défaut pour le calcul

Le premier tableau obtenu est relatif aux observations. Pour chaque concentration horaire observée sur la période comprise entre les deux seuils fixés, le tableau indique le nom de la station concernée, la date, la valeur de la concentration horaire observée, la concentration horaire simulée au même instant (case grise si absence de calcul), et le biais entre ces deux valeurs selon la formule : simulation - observation.

Le second tableau est relatif aux simulations. De la même façon, pour chaque concentration horaire simulée sur la période comprise entre les deux seuils fixés, le tableau indique le nom de la station concernée, la date, la valeur de la concentration horaire simulée, la concentration horaire observée au même instant (case grise si donnée invalide), et le biais entre ces deux valeurs selon la formule : simulation - observation.

Pour chacun des deux tableaux, un biais moyen est calculé sur l'ensemble des heures selon la formule suivante :

$$Biais\ moyen = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)$$

Où O_i et S_i représentent respectivement une concentration horaire observée et simulée, et N le nombre total de couples (O_i, S_i) utilisés. **Notons que par construction, pour le calcul de ce paramètre, seules les heures de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.**

3.4 Case « Tableaux contingences seuil à choisir »

Le cadre suivant affiche le tableau de contingence relatif à la bonne prévision d'un dépassement horaire d'un seuil sur la période, ce seuil exprimé en µg/m³ pouvant être rentré à la main par l'utilisateur. Par défaut, le seuil de concentration est fixé à 80 µg/m³ pour PM, 120 µg/m³ pour O_3, 200 µg/m³ pour NO_2 et 300 µg/m³ pour SO_2. **Notons que pour le cas particulier de l'O_3, des calculs seront aussi réalisés pour le dépassement du seuil 120 µg/m³ en moyenne glissante sur 8h. Afin de ne pas confondre ce seuil avec celui associé aux statistiques de dépassements horaires du seuil 120 µg/m³, nous le renommerons, par exemple par « 120HG ».**

Le tableau de contingence affiché est de la forme :

Nb échéances horaires : N	**Obs ≥ seuil µg/m³ : n1**	**Obs < seuil µg/m³ : n2**
Prévision ≥ seuil µg/m³ : n3	A	B
Prévision < seuil µg/m³ : n4	C	D

Avec A le nombre d'heures sur la période avec Observation ≥ seuil et Prévision ≥ seuil (nombre d'alertes correctes), B le nombre d'heures sur la période avec Observation < seuil et Prévision ≥ seuil (nombre de fausses alertes), C le nombre d'heures sur la période avec Observation ≥ seuil et Prévision < seuil (nombre d'événements manqués) et D le nombre d'heures sur la période avec Observation < seuil et Prévision < seuil (nombre de non alertes correctes). N représente le nombre de couples (observation horaire, simulation horaire) disponibles sur la période sélectionnée, n1 le nombre d'heures observées qui dépassent le seuil, n2 le nombre d'heures observées qui ne dépassent pas le seuil, n3 le nombre d'heures simulées qui dépassent le seuil et n4 le nombre d'heures simulées qui ne dépassent pas le seuil. **Notons que ce tableau de contingence est réalisé uniquement sur les N heures de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant.**

Différents scores sont associés à ce tableau[2] :

- Taux de bonne prévision (%) : il indique le pourcentage d'alertes ou non-alertes bien prévues et se calcule par la formule $100 \times \frac{A+D}{A+B+C+D}$
- Taux de prévision fausse (%) : il indique le pourcentage de fausses alertes ou alertes manquées et se calcule par la formule $100 \times \frac{B+C}{A+B+C+D}$
- Paramètre Alpha : il indique la probabilité de détection (POD) et se calcule par la formule $\frac{A}{A+C}$
- Paramètre Bêta : il indique le taux de fausses alertes (FAR) et se calcule par la formule $\frac{B}{A+B}$
- Moyenne obs≥seuil µg/m³ : il indique la moyenne des concentrations horaires mesurées dépassant le seuil.

[2] *Une brève description de la signification de ces scores sera indiquée sur SuiviStat par des « infos bulles ».*

- Moyenne sim≥seuil µg/m³ : il indique la moyenne des concentrations horaires simulées dépassant le seuil, en ne considérant que les échéances horaires pour lesquelles une observation existe au même instant.
- Biais moyen (µg/m³) : il s'agit de la différence entre les deux valeurs précédentes selon la formule « Moyenne sim – Moyenne obs ».
  Notons que pour le calcul de ces paramètres, seules les N heures de la période étudiée pour lesquelles observations et simulations sont disponibles au même instant seront considérées.
- Moyenne sim≥seuil µg/m³ : il indique la moyenne des concentrations horaires simulées dépassant le seuil, en considérant toutes les échéances horaires, y compris celles pour lesquelles aucune observation n'est présente au même instant. **Ainsi pour ce dernier paramètre, l'ensemble des simulations disponibles sont utilisées, y compris celles pour lesquelles aucune observation correspondante n'existe.**

3.5 Case « Tableaux AOT pour O_3 seuil à choisir »

Comme son nom l'indique, cette option est valable uniquement dans le cas de la plateforme O_3. Il est possible de paramétrer le calcul en rentrant à la main un seuil de concentration exprimé en ppb[3]. Par défaut, le seuil de concentration est fixé à 40 ppb. Le tableau obtenu fournit différents indices :

- Cumul des obs ≥ 40 ppb/heures : cette valeur consiste à faire la somme des observations horaires ayant dépassé le seuil fixé entre 8h et 22h au cours de la période.
- Cumul des sim ≥ 40 ppb/heures : cette valeur consiste à faire la somme des simulations horaires ayant dépassé le seuil fixé entre 8h et 22h au cours de la période, en ne considérant que les échéances horaires pour lesquelles une observation existe au même instant.
- Biais (ppb/heures) : il s'agit de la différence entre les deux valeurs précédentes selon la formule : cumul des sim – cumul des obs.
- Cumul des sim ≥ 40 ppb/heures : cette valeur consiste à faire la somme des simulations horaires ayant dépassé le seuil fixé entre 8h et 22h au cours de la période, en considérant toutes les échéances horaires, y compris celles pour lesquelles aucune observation n'est présente au même instant.

[3] *Facteur de conversion : 1 ppb = 2 µg/m³ pour O_3*

4 Format des fichiers statistiques utilisés pour les calculs de performance et format des statistiques à afficher sur SuiviStat

4.1 Format des fichiers statistiques utilisés pour les calculs de performance

Répertoire de stockage du fichier statistique

Chaque jour, un répertoire est créé par chaque plateforme (O_3, PM, SO_2, NO_2) afin de stocker le fichier statistiques (fichier "stat.txt") contenant les observations et simulations horaires réalisées pour la veille. Le répertoire ainsi créé porte le nom de la date de sa création, au format *aaaammjj*. Cependant, les observations et simulations inscrites dans le fichier stat qu'il contient sont relatives au jour précédent. Ainsi, par exemple, le répertoire dénommé "20080420", créé par une application le 20/04/2008, contiendra le fichier stat des observations et simulations pour la journée du 19/04/2008. Aussi, **il est indispensable de tenir compte de ce décalage lors de la lecture des fichiers stat** nécessaires aux calculs de statistiques, qu'elles soient journalières ou calculées sur une plus longue période. Par exemple, pour évaluer les performances d'une application pour la journée du 19/04/2008, le fichier stat à prendre en compte sera celui sauvegardé dans le répertoire dénommé "20080420".

Contenu du fichier statistique

Un fichier stat est un fichier texte dénommé "stat.txt" créé chaque jour et contenant pour chaque heure de la journée et pour chaque station la valeur de concentration mesurée pour le polluant, ainsi que les valeurs simulées par les différents cycles de simulation (J-1, J, J+1, et J+2 en fonction de la plateforme considérée) et selon les différentes méthodes de calcul (simulation brute, simulation brute+biais, etc.). Une valeur de concentration mesurée ou simulée manquante est identifiée par les codes "-999" et "-9999". Le nom des colonnes relatives aux simulations informe sur le cycle de simulation et l'étape de calcul considérés. Bien que les fichiers stat possèdent une colonne indiquant une étape de calcul analysée pour tous les cycles de simulation (J-1, J, J+1 et J+2), celle-ci n'est en réalité renseignée que pour le cycle de simulation J-1. Les colonnes relatives aux simulations varient entre les applications O_3, PM, SO_2, NO_2. Les colonnes des fichiers stat pour les polluants O_3, PM, et NO_2 sont présentées dans le Tableau 7.

O_3	PM	NO_2
dates	dates	dates
stations	stations	stations
mesures	mesures	mesures
J-1 PREVAIR	J-1 ADMS	J-1 PREVAIR
J-1 PREVAIR+biais	J-1 ADMS+fond	J-1 PREVAIR+dérivé externe
J-1 PREVAIR+biais analysé	J-1 ADMS+fond+biais	J-1 PREVAIR+dérivé externe analysée
J PREVAIR	J-1 ADMS+fond+biais analysé	J PREVAIR
J PREVAIR+biais	J ADMS	J PREVAIR+dérivé externe
J PREVAIR+biais analysé	J ADMS+fond	J PREVAIR+dérivé externe analysée
J+1 PREVAIR	J ADMS+fond+biais	J+1 PREVAIR
J+1 PREVAIR+biais	J ADMS+fond+biais analysé	J+1 PREVAIR+dérivé externe
J+1 PREVAIR+biais analysé	J+1 ADMS	J+1 PREVAIR+dérivé externe analysée
J+2 PREVAIR	J+1 ADMS+fond	J+2 PREVAIR
J+2 PREVAIR+biais	J+1 ADMS+fond+biais	J+2 PREVAIR+dérivé externe
J+2 PREVAIR+biais analysé	J+1 ADMS+fond+biais analysé	J+2 PREVAIR+dérivé externe analysée

Tableau 7 : Colonnes présentes dans les fichiers statistiques des polluants O_3, PM et NO_2

Dans le cas du SO_2, les simulations de la pollution sont également présentées par industriel de la zone AIRFOBEP (ESSO, TOTAL…) (voir Tableau 8). Cependant, pour les calculs statistiques, il ne sera

tenu compte que des colonnes portant sur la simulation correspondant à la somme de la pollution de tous les industriels (colonnes notées en caractères gras dans le Tableau 8).

SO_2
dates
stations
mesures
J-1 TOUS ADMS
J-1 TOUS ADMS+correction
J-1 TOUS ADMS+correction assimilé
J-1 BP ADMS
J-1 BP ADMS+correction
J-1 EDF ADMS
J-1 EDF ADMS+correction
J-1 ESSO ADMS
J-1 ESSO ADMS+correction
J-1 LAFARGE ADMS
J-1 LAFARGE ADMS+correction
J-1 NAPHTACHIMIE ADMS
J-1 NAPHTACHIMIE ADMS+correction
J-1 SBR ADMS
J-1 SBR ADMS+correction
J-1 SHELL CHIMIE ADMS
J-1 SHELL CHIMIE ADMS+correction
J-1 SNET ADMS
J-1 SNET ADMS+correction
J-1 SOLLAC ADMS
J-1 SOLLAC ADMS+correction
J-1 TOTAL ADMS
J-1 TOTAL ADMS+correction
J TOUS ADMS
J BP ADMS
J EDF ADMS
J ESSO ADMS
J LAFARGE ADMS
J NAPHTACHIMIE ADMS
J SBR ADMS
J SHELL CHIMIE ADMS
J SNET ADMS
J SOLLAC ADMS
J TOTAL ADMS
J+1 TOUS ADMS
J+1 BP ADMS
J+1 EDF ADMS
J+1 ESSO ADMS
J+1 LAFARGE ADMS
J+1 NAPHTACHIMIE ADMS
J+1 SBR ADMS
J+1 SHELL CHIMIE ADMS
J+1 SNET ADMS
J+1 SOLLAC ADMS
J+1 TOTAL ADMS
J+2 TOUS ADMS
J+2 BP ADMS
J+2 EDF ADMS

J+2 ESSO ADMS
J+2 LAFARGE ADMS
J+2 NAPHTACHIMIE ADMS
J+2 SBR ADMS
J+2 SHELL CHIMIE ADMS
J+2 SNET ADMS
J+2 SOLLAC ADMS
J+2 TOTAL ADMS

Tableau 8 : Colonnes présentes dans les fichiers statistiques du polluant SO_2 (les colonnes en gras sont celles à utiliser pour les calculs statistiques)

4.2 Format des statistiques à afficher sur SuiviStat

Pour l'ensemble des statistiques exprimées en µg/m³ (concentration moyenne mesurée, concentration moyenne simulée, biais moyen, écart-type du biais élargi à 2, quantile 0.025, quantile 0.975), une précision de 2 chiffres après la virgule est requise. Pour les statistiques exprimées en pourcentage (NMSE, E-20%, bonnes prévisions, prévisions sous-estimantes, prévisions sur-estimantes pour l'indice ATMO), ces chiffres seront arrondis à l'entier le plus proche. Enfin pour les autres paramètres (corrélation R^2, biais moyen sur l'indice ATMO), une précision de 2 chiffres après la virgule est requise.

4.3 Affichage sur SuiviStat des dates non prises en compte dans les calculs

Dans le cas où l'on cherche à réaliser des calculs statistiques sur une période comprenant des journées pour lesquelles on ne dispose pas de fichier stat, les journées non prises en compte dans les calculs doivent être signalées sur SuiviStat grâce à un message en bas de page comme présenté dans la Figure 3.

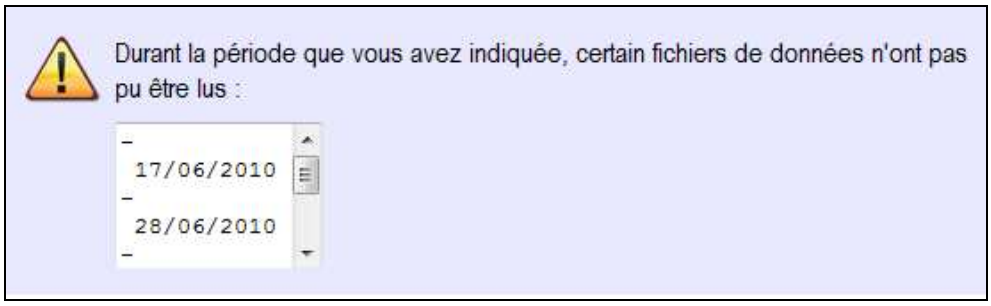


Figure 3 : Message à afficher dans le cas de journées avec absence de fichier statistique

5 Cas particulier de la plateforme IQA

5.1 Calculs statistiques journaliers

La visualisation prévue sous SuiviStat des performances de la plateforme IQA pour un jour particulier est montrée dans la Figure 4.

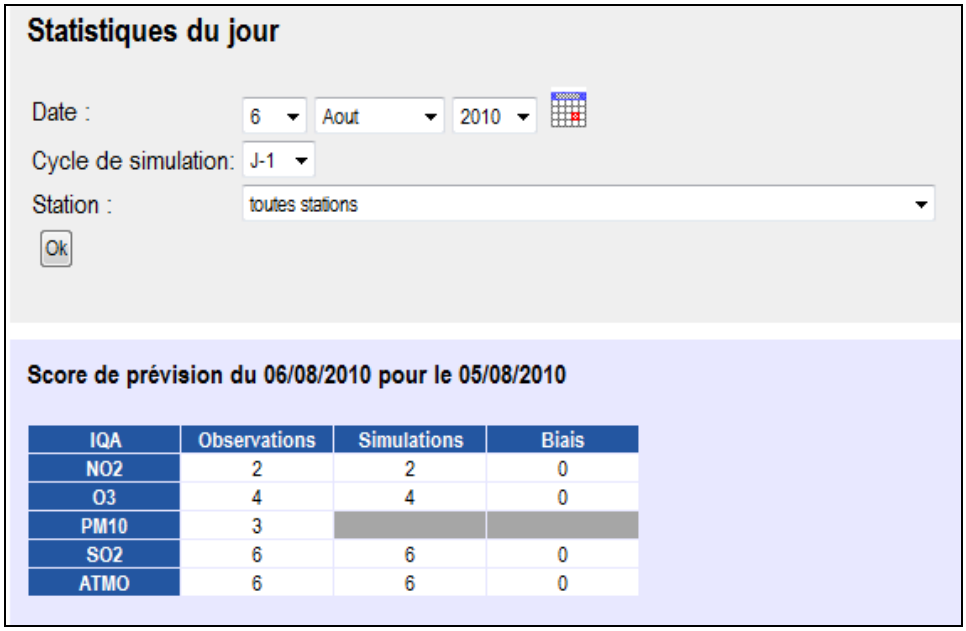


Figure 4 : Interface type prévue pour l'affichage de scores de performances journaliers pour la plateforme ATMO.

Comme pour les autres plateformes, le cadre supérieur (« Statistiques du jour ») permet de définir différents paramètres :

- Date : **il s'agit du jour d'observation que l'on désire étudier**. Par défaut, la date affichée est donc celle **de la veille** (dernier jour pour lequel les observations sont disponibles).
- Cycle de simulation : il sera ici possible de choisir si l'on désire évaluer les performances de l'application en mode « prévision » (J, J+1 et J+2 en fonction de la plate-forme considérée) ou « analyse » (calculs effectués pour J-1 en tenant compte de données d'observations pour la météorologie et la pollution).
- Station : il est possible de réaliser des calculs de performances en considérant une seule station du réseau AIRFOBEP, toutes les stations à la fois, ou les stations appartenant à une zone prédéfinie.

Ici, on ne dispose plus du paramètre « Méthode de calcul », comme pour les autres plateformes. En effet, les statistiques sont ici réalisées à partir de fichiers .pgm, dans lesquels sont uniquement stockées les valeurs de la dernière étape de calcul pour chaque cycle de prévision. Ainsi, pour J-1, les statistiques portant sur l'indice ATMO seront celles en mode « analyse » avec prise en compte de l'ensemble des observations disponibles.

Le Tableau 9 récapitule les différentes valeurs pouvant être prises par chaque paramètre.

	Indice ATMO
Date	Consiste simplement à choisir le jour, le mois et l'année d'intérêt
Cycle de simulation	J-1, J, J+1 et J+2
Station	- ARLS, MRMV, SMMR, SRMY, RBRT, VTRL, BETG, BMGS, ISTR, FSCB, FSMR, MRGV, MEDE, CHNF, MILE, MNDM, SSLP, MPTI, PDBC, MLVR, PDBL, PDBE, PSLV, SLPV, FOLV, CLRT - Zone n°1 (Arles/St-Martin-de-Crau) : ARLS, MRMV, SMMR, SRMY - Zone n°2 (Berre l'Etang) : BETG, BMGS, RBRT, VTRL - Zone n°3 (Fos-sur-mer) : FSMR, FSCB, ISTR - Zone n°4 (Istres/Miramas/St-Chamas) : ISTR, MRMV, FSCB - Zone n°5 (Marignane/St-Victoret/Chateauneuf-les-Martigues/Gignlas-la-Nerthe) : MRGV, CHNF, MEDE, VTRL, RBRT - Zone n°6 (Martigues ville/St-Mitre-les-Remparts) : MILE, MPTI, MNDM, PDBC, MLVR, SSLP - Zone n°7 (Port-de-Bouc) : PDBL, PDBC, MILE, FSCB, ISTR - Zone n°8 (Port-St-Louis) : PSLV, ISTR, FSCB, CRAU - Zone n°9 (Salon de Provence/Cornillon Confoux) : SLPV, FOLV, MRMV - Zone n°10 (Carry-le-Rouet/Sausset-les-Pins) : SSLP, CLRT, MILE, MNDM - Zone n°11 (Vitrolles/Rognac/Coudoux/Velaux/Ventrabren) : VTRL, RBRT - Toutes stations

Tableau 9 : Valeurs prises par les différents paramètres de configuration pour la plateforme ATMO pour un calcul de performances journalières.

Un tableau permet ensuite d'obtenir les valeurs de sous-indices ATMO pour les polluants O_3, PM, SO_2 et NO_2 :

- Observations : il s'agit de la valeur de sous-indice ATMO associée à l'indice de la concentration du polluant observée lors de la journée.
  - Si une seule station est sélectionnée, il s'agira du sous-indice associé à la concentration (concentration moyenne journalière pour les PM et concentration maximale journalière pour O_3, SO_2 et NO_2) observée à la station.
  - Si le calcul porte sur toutes les stations ou une zone, il s'agira du sous-indice associé à la moyenne de la concentration (concentration moyenne journalière pour les PM et concentration maximale journalière pour O_3, SO_2 et NO_2) observée en chaque station.
- Simulations : il s'agit de la valeur de sous-indice ATMO associée à l'indice de la concentration du polluant simulée lors de la journée
  - Si une seule station est sélectionnée, il s'agira du sous-indice associé à la concentration (concentration moyenne journalière pour les PM et concentration maximale journalière pour O_3, SO_2 et NO_2) simulée à la station.
  - Si le calcul porte sur toutes les stations ou une zone, il s'agira du sous-indice associé à la moyenne de la concentration (concentration moyenne journalière pour les PM et concentration maximale journalière pour O_3, SO_2 et NO_2) simulée en chaque station.
- Biais : il s'agit de la différence entre les valeurs "Simulations" et "Observations" selon la formule : Simulations - Observations.

La dernière ligne indique concerne l'indice ATMO finalement obtenu au cours de la journée :

- Observations : il s'agit du maximum des quatre valeurs de sous-indices ATMO observées.
- Simulations : il s'agit du maximum des quatre valeurs de sous-indices ATMO simulées.
- Biais : il s'agit de la différence entre les valeurs "Simulations" et "Observations" selon la formule : Simulations - Observations.

5.2 Calculs statistiques sur une période

La visualisation prévue sous SuiviStat des performances de la plateforme IQA sur une période est montrée dans la Figure 5.

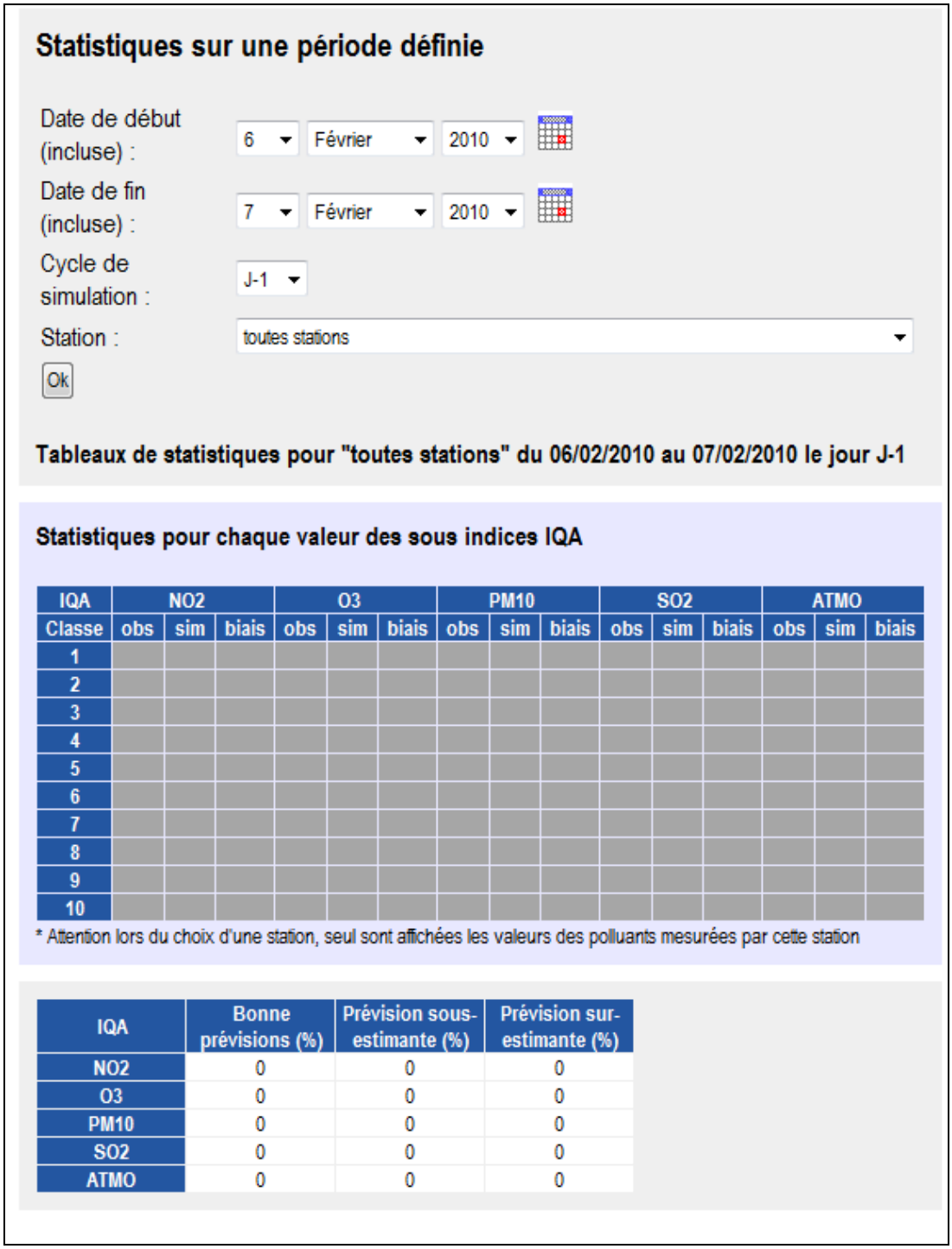


Figure 5 : Interface type prévue pour l'affichage de scores de performances sur une période pour la plateforme ATMO.

Comme pour les autres plateformes, le cadre supérieur (« Statistiques du jour ») permet de définir différents paramètres :

- Date de début (incluse) : il s'agit du premier jour **d'observation** (inclus) de la période que l'on désire étudier.
- Date de fin (incluse) : il s'agit du dernier jour **d'observation** (inclus) de la période que l'on désire étudier
- Cycle de simulation (voir Tableau 9) : il sera ici possible de choisir si l'on désire évaluer les performances de l'application en mode « prévision » (J, J+1 et J+2 en fonction de la plate-forme considérée) ou « analyse » (calculs effectués pour J-1 en tenant compte de données d'observations pour la météorologie et la pollution).
- Station (voir Tableau 9) : il est possible de réaliser des calculs de performances en considérant une seule station du réseau AIRFOBEP, toutes les stations à la fois, ou les stations appartenant à une zone prédéfinie.

Le premier tableau obtenu présente, pour chaque polluant et chaque valeur de sous-indice ATMO ou indice ATMO observé, les paramètres suivants :

- Obs : le nombre de sous-indices ATMO observés sur la période considérée. **Ici, seules les données d'observations ATMO pour lesquelles les simulations ATMO correspondantes sont disponibles seront utilisées.**
- Sim : le nombre de sous-indices ATMO simulés sur la période considérée. **Ici, seules les données de simulations ATMO pour lesquelles les observations ATMO correspondantes sont disponibles seront utilisées.**
- Biais : il s'agit de l'erreur commise pour la valeur de sous-indice ATMO observé, selon la formule :

$$Biais\ moyen = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)$$

Où N représente le nombre de couples (O_i, S_i) utilisés, et O_i et S_i représentent respectivement une valeur d'indice ATMO observé et simulé. **Notons que par construction, pour le calcul de ce paramètre, seules les journées de la période étudiée pour lesquelles observations ATMO et simulations ATMO sont disponibles au même instant seront considérées.**

Un tableau permet ensuite d'obtenir différents critères concernant les sous-indices ATMO des quatre polluants et l'indice ATMO final :

- Bonne prévision (%) : il s'agit du nombre de biais ATMO égaux à 0 rapporté au nombre de journées pour lesquelles le biais ATMO est disponible.

$$Bonnes\ prévisions = 100 \times \frac{1}{N}\sum_{i=1}^{N} previ_i \quad \text{avec } previ_i = 1 \text{ si } Biais\ ATMO_i = 0$$

$$previ_i = 0 \text{ sinon.}$$

Où N représente le nombre de journées utilisées (i.e auxquelles le biais ATMO était disponible), et $previ_i$ est un indicateur permettant de comptabiliser le nombre de biais ATMO égaux à 0.

- Prévision sous-estimante (%) : il s'agit du nombre de biais ATMO négatifs rapporté au nombre de journées pour lesquelles le biais ATMO est disponible
- Prévision sur-estimante (%) : il s'agit du nombre de biais ATMO positifs rapporté au nombre de journées pour lesquelles le biais ATMO est disponible.

Notons que par construction, pour le calcul de ces paramètres, seules les journées de la période étudiée pour lesquelles observations ATMO et simulations ATMO sont disponibles au même instant seront considérées.