

# Actualisation de la classification des sites de mesure

Mathieu Joly & Valentin Petiot

23 janvier 2024

## Table des matières

<b>1</b>	<b>Données utilisées</b>	<b>2</b>
<b>2</b>	<b>Traitement des données</b>	<b>2</b>
<b>3</b>	<b>Cartographie du résultat</b>	<b>3</b>
<b>4</b>	<b>Validation croisée</b>	<b>3</b>
<b>5</b>	<b>Étude des anomalies</b>	<b>3</b>
<b>6</b>	<b>Comparaison à la précédente version</b>	<b>3</b>
<b>7</b>	<b>Évolution du jeu de stations classifiées</b>	<b>12</b>
<b>8</b>	<b>Conclusion</b>	<b>12</b>

## 1 Données utilisées

- La période d'étude comprend 8 années, de 2016 à 2023, avec des données non validées pour 2022 et 2023.
- Ne sont pas pris en compte les sites d'altitude supérieure à 1400 m (altitude à partir de laquelle le nombre de stations diminue fortement). En Europe, ces stations sont peu nombreuses, mais ne peuvent pas être confondues avec les sites de plaine pour l'analyse.
- Les stations renseignées comme « industrielles » ne sont pas prises en compte. La variabilité temporelle de ce type de mesure est très difficile à caractériser, et la méthode n'est pas suffisamment robuste pour appréhender le comportement potentiellement erratique des indicateurs calculés.
- Désormais, les mesures de CO à très faibles résolution (valeurs discrètes multiples de 100) sont éliminées en amont. Ces stations ne sont de toutes façons pas utilisées en aval par CAMS2\_40.

À partir des métadonnées, on dérive la typologie simplifiée suivante :

**Type R** : sites qualifiés *background* et *rural*.

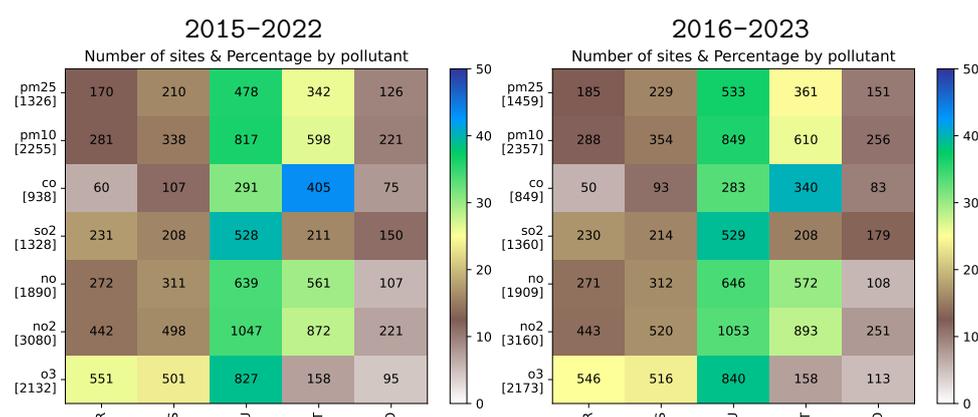
**Type S** : sites qualifiés *background* et *suburban*.

**Type U** : sites qualifiés *background* et *urban*.

**Type T** : sites qualifiés *traffic* et *urban*.

**Type O** : toutes les autres stations, ainsi que les stations en dehors du sous-domaine, qui ne seront pas prises en compte pour l'Analyse Discriminante, mais qui seront classifiées *a posteriori*.

Le CO et l'ozone font toujours figure d'anomalie, avec beaucoup de stations T dans le premier cas, et beaucoup de stations *background* (R, S, et U) dans l'autre. Le réseau de mesure s'étoffe doucement, en particulier pour les PM<sub>2.5</sub> (+10%) et les PM<sub>10</sub> (+5%).



**Figure 1** – Nombre de stations sélectionnées (données suffisantes), par type de métadonnée. Les couleurs correspondent au pourcentage par polluant.

## 2 Traitement des données

Les figures 2 et 3 montrent que les séries temporelles sur certaines régions, bien que suffisantes en quantité de données, ne permettent pas de calculer tous les indicateurs. C'est le cas pour certaines régions d'Italie, ainsi qu'en Allemagne, Tchéquie, et Grande Bretagne pour le

NO et le SO<sub>2</sub>. Pour ces stations, les valeurs absentes sont trop nombreuses au sein de chaque journée, ce qui empêche le calcul d'indicateurs quotidiens. Pour l'Allemagne, il semble que les faibles valeurs, autrefois mises à zéro, aient été supprimées de la base de données, d'où un taux de valeurs absentes trop élevé. Les autres pays n'ont pas procédé à un tel nettoyage des valeurs proches de zéro.

### 3 Cartographie du résultat

Les figures 4 et 5 illustrent la classification obtenue pour chaque polluant.

### 4 Validation croisée

La figure 6 compare les « validations croisées » par rapport aux types dérivés des métadonnées. La cohérence entre les classifications subjective (métadonnées) et objective est stable par rapport à la précédente version.

### 5 Étude des anomalies

Nous allons nous intéresser aux comportements marginaux de la figure 6 :

- le pourcentage des stations R qui se retrouvent dans les classes 6-10.
- le pourcentage des stations S, U et T qui se retrouvent dans les classes 1-3.

	O <sub>3</sub>	NO <sub>2</sub>	NO	SO <sub>2</sub>	CO	PM <sub>10</sub>	PM <sub>2.5</sub>
R 6-10	4 → 4	2 → 2	4 → 4	30 → 28	16 → 18	20 → 19	28 → 24
S+U+T 1-3	8 → 9	3 → 3	4 → 4	13 → 13	4 → 3	6 → 7	9 → 8

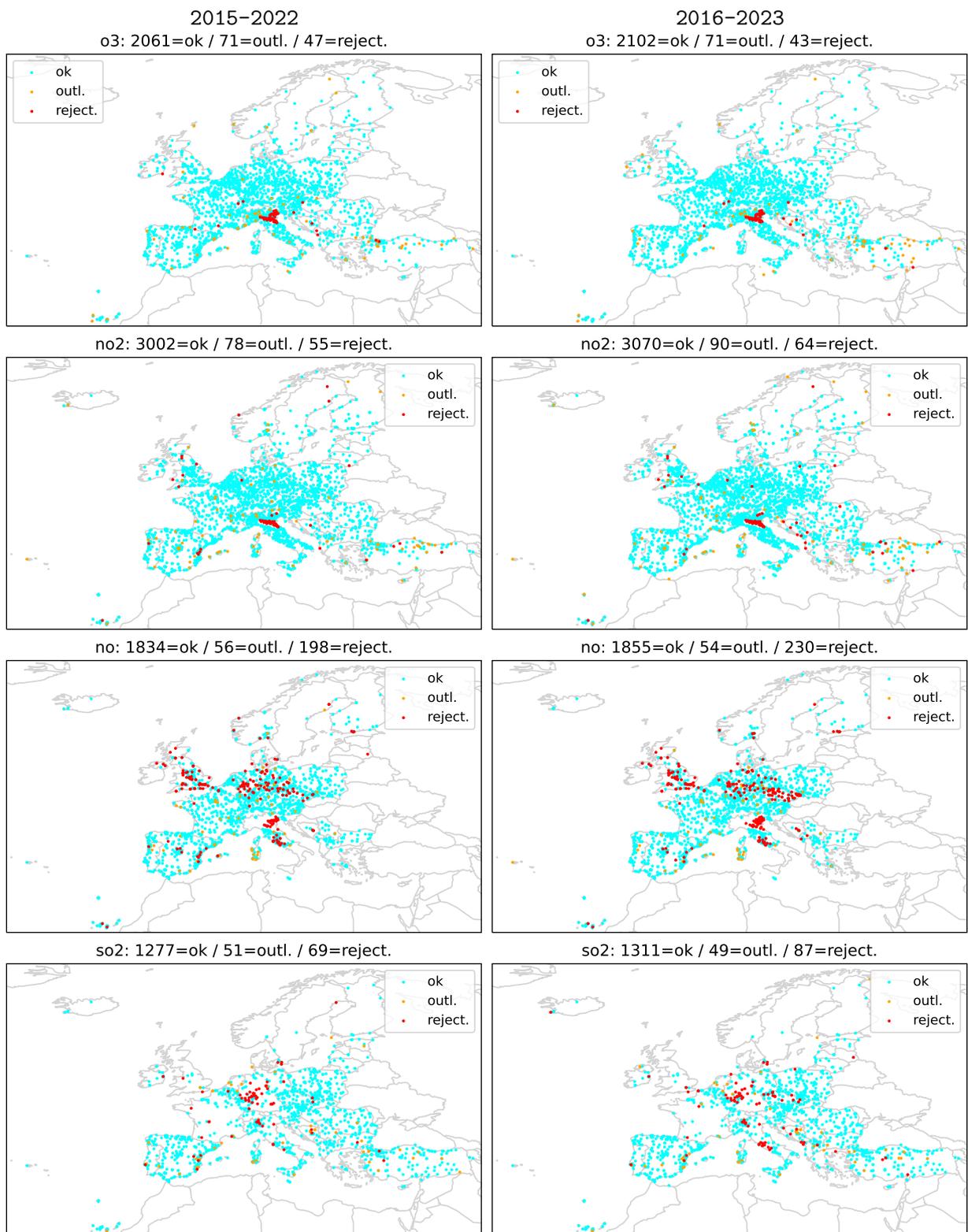
**Tableau 1** – Pourcentage des anomalies (cf. paragraphe ci-dessus). Évolution entre l'ancienne et la nouvelle classification (en vert pour une amélioration, en rouge pour une détérioration, et surligné de jaune quand plus de 2% des stations sont affectées).

Le tableau 1 montre que pour les stations rurales, la classification comporte des anomalies moins nombreuses que l'année précédente pour les PM<sub>2.5</sub>. Pour les autres polluants, les différences sont faibles.

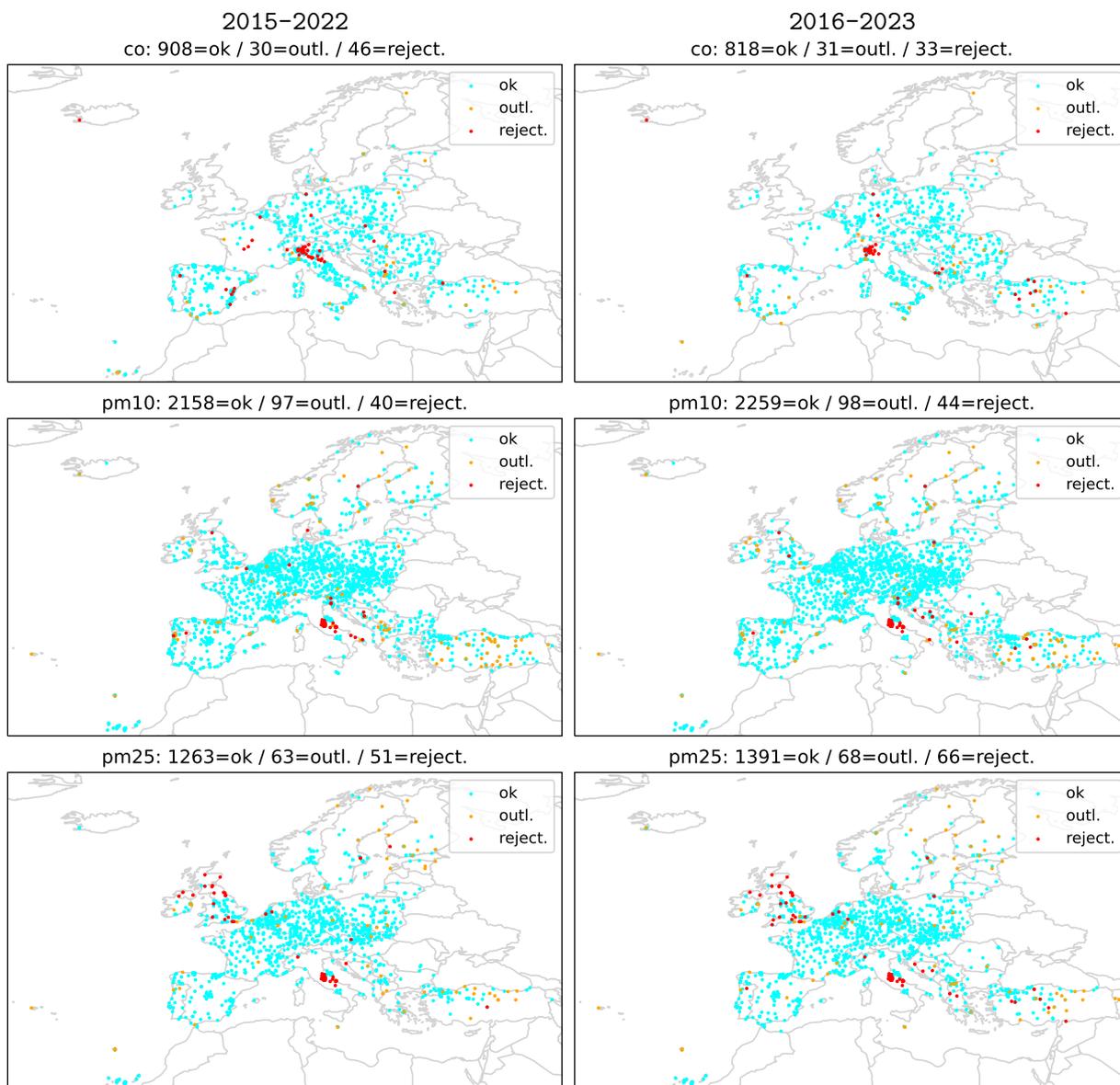
Les cartes 7 et 8 cartographient les anomalies du tableau 1. L'analyse est difficile, car il faudrait regarder localement la configuration de chacun de ces sites « douteux », et les sources de pollution environnantes. La nouvelle version ne modifie pas beaucoup la localisation de ces anomalies.

### 6 Comparaison à la précédente version

Pour les stations en commun dans les deux classifications, la figure 9 compare les classes obtenues. La classification est généralement très stable, à l'exception de quelques anomalies isolées. C'est comme à chaque fois pour les PM<sub>2.5</sub> que la cohérence est la moins bonne entre les deux versions.



**Figure 2** – Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.



**Figure 3** – Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

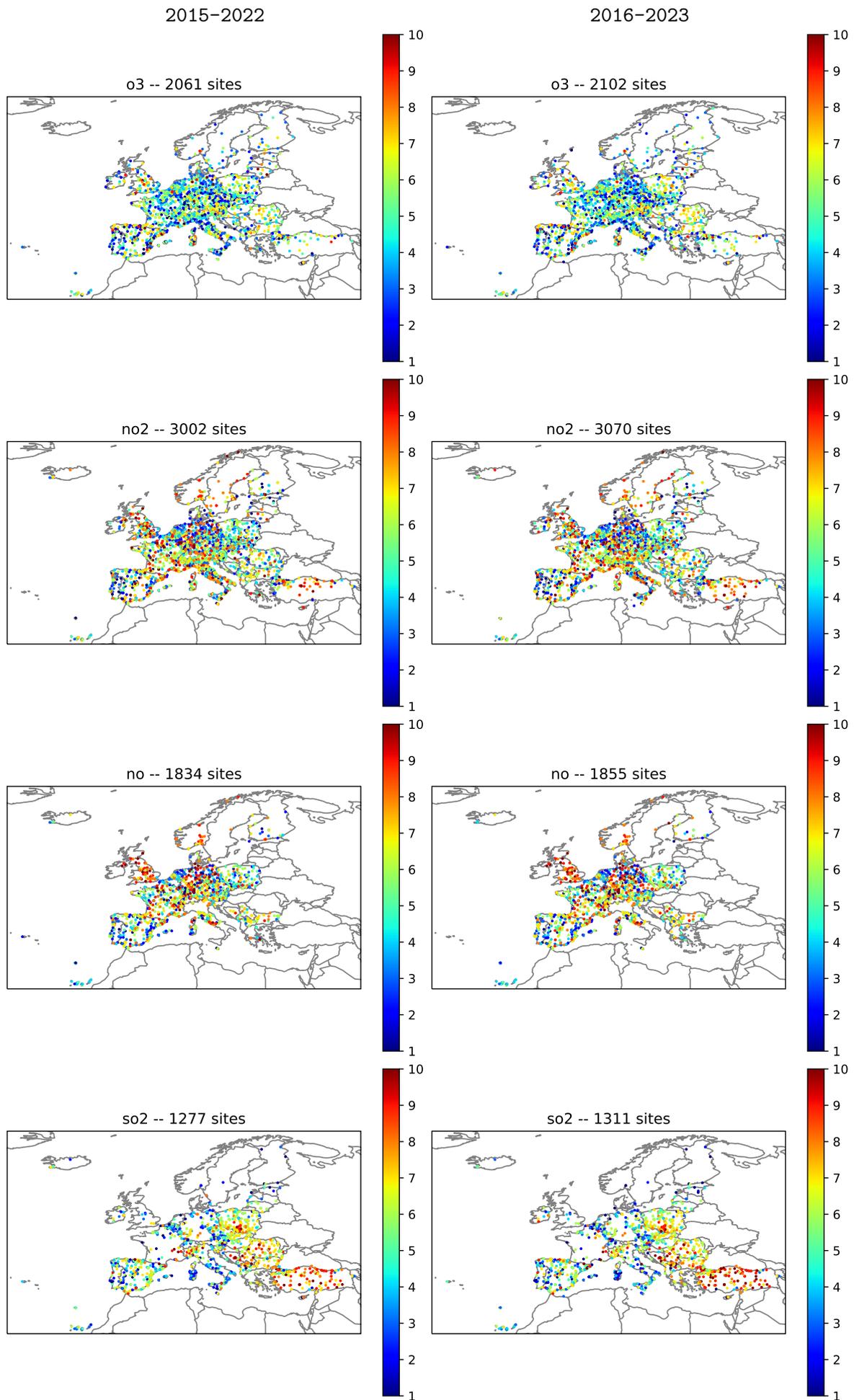
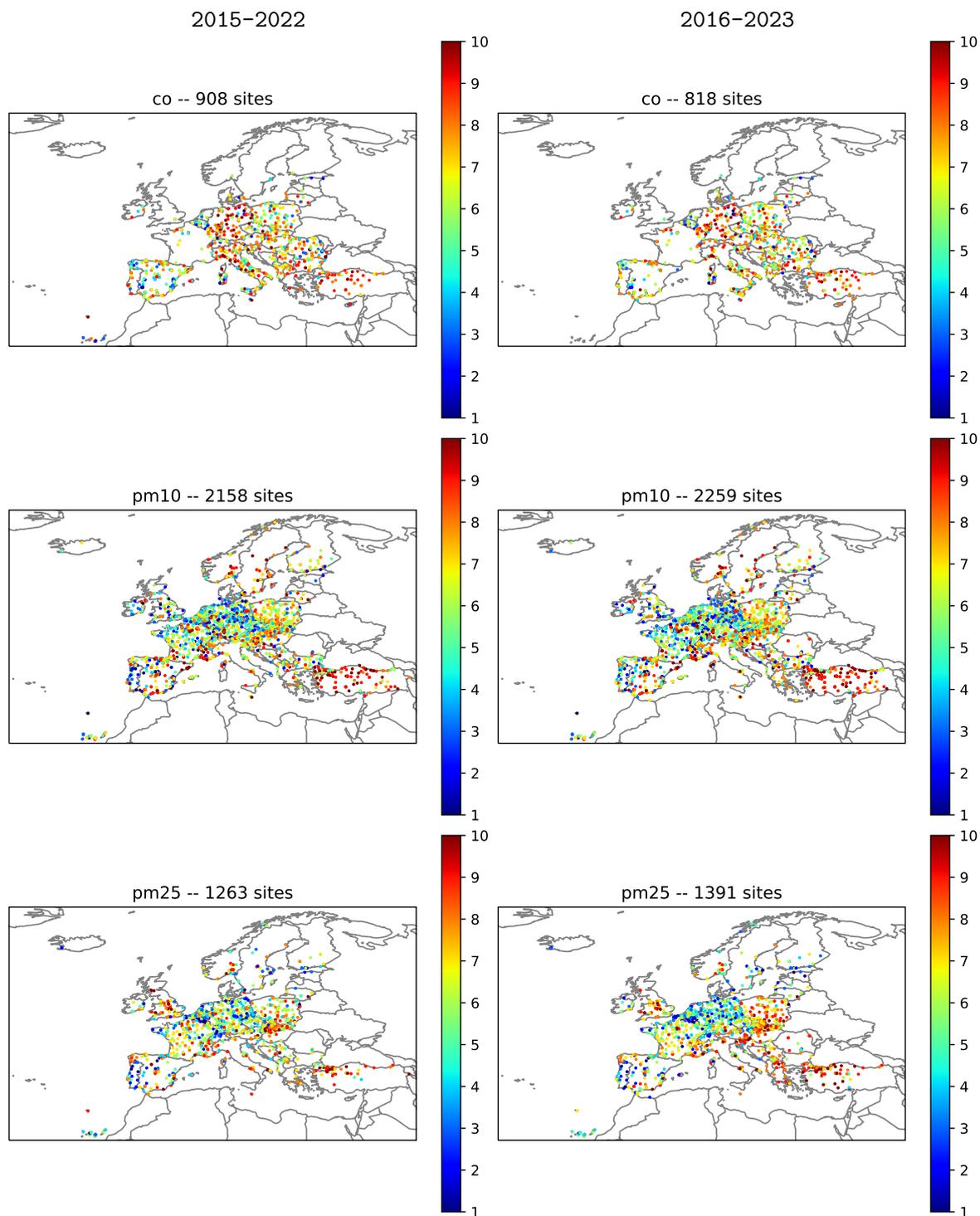


Figure 4 – Cartographie de la classification obtenue.



**Figure 5** – Cartographie de la classification obtenue. À gauche, pour la précédente classification; et à droite, pour la nouvelle version.

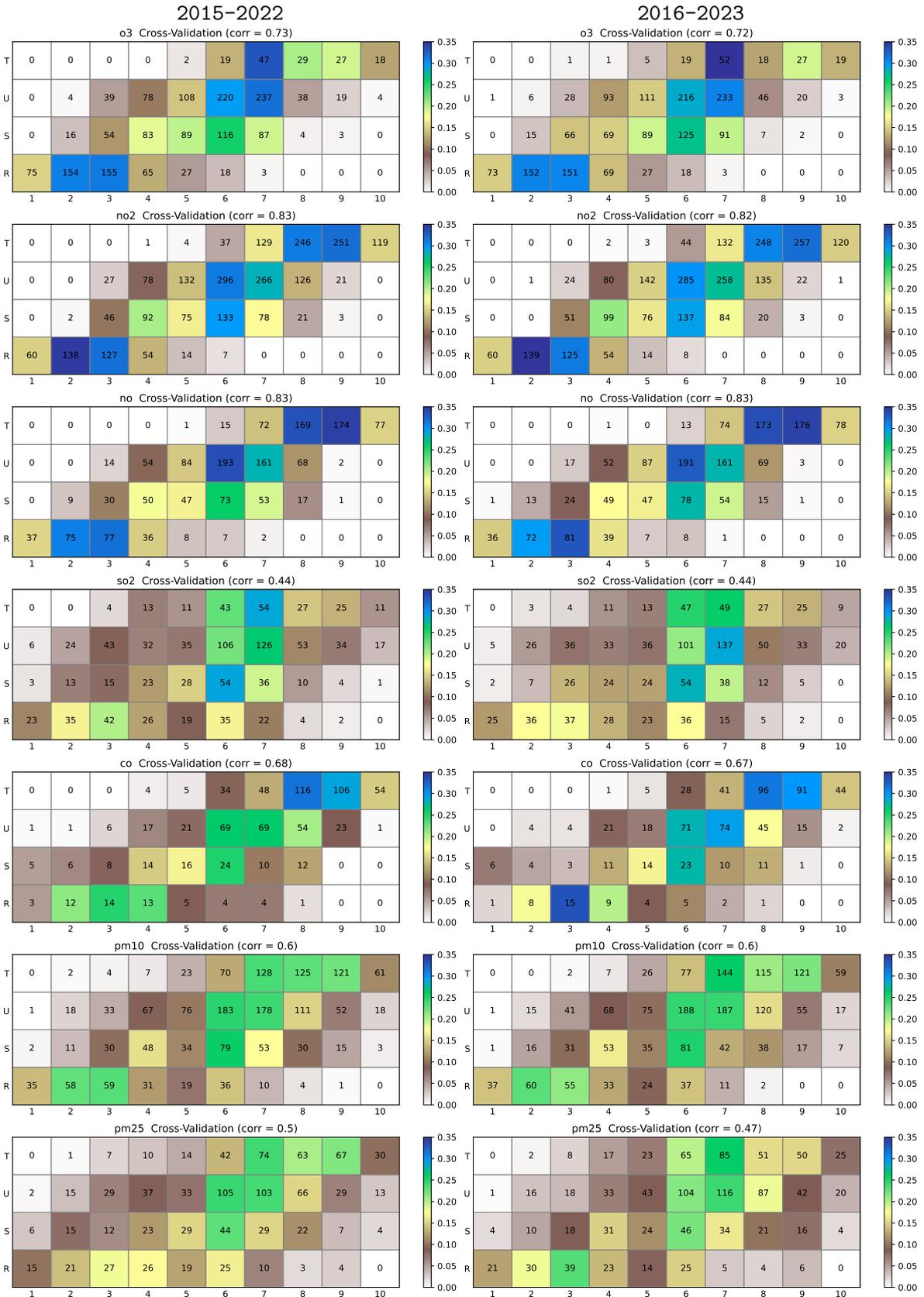
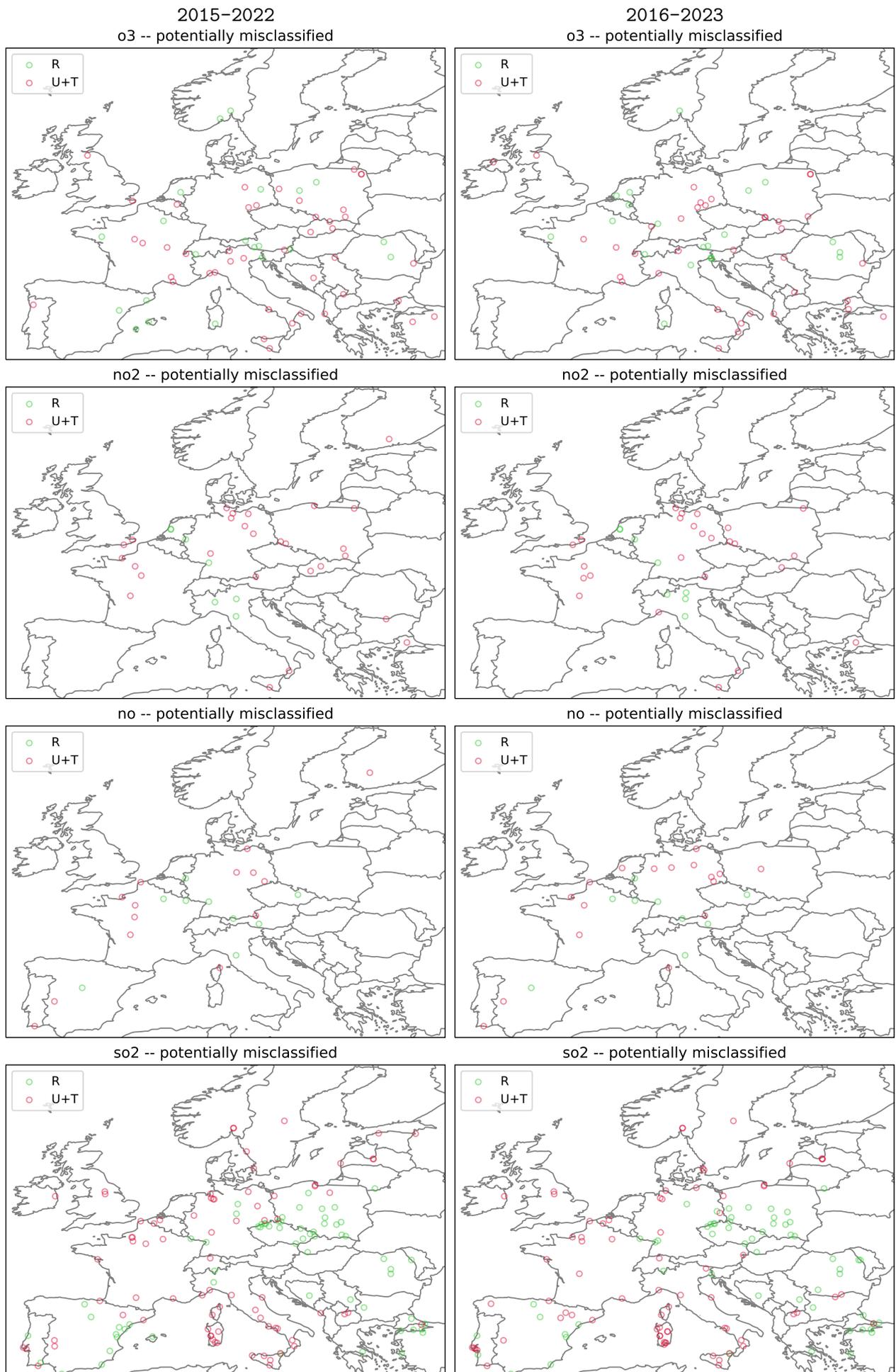
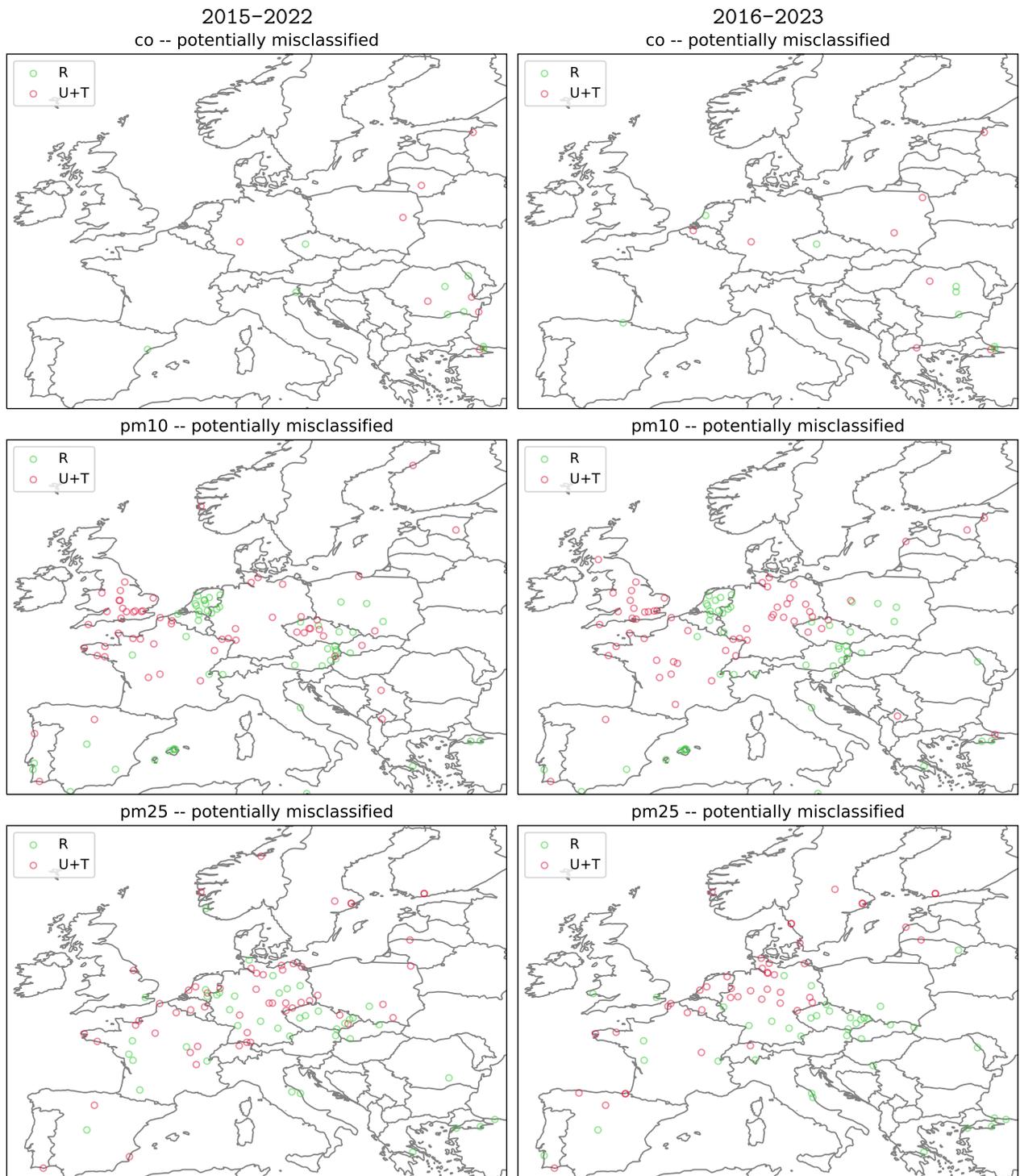


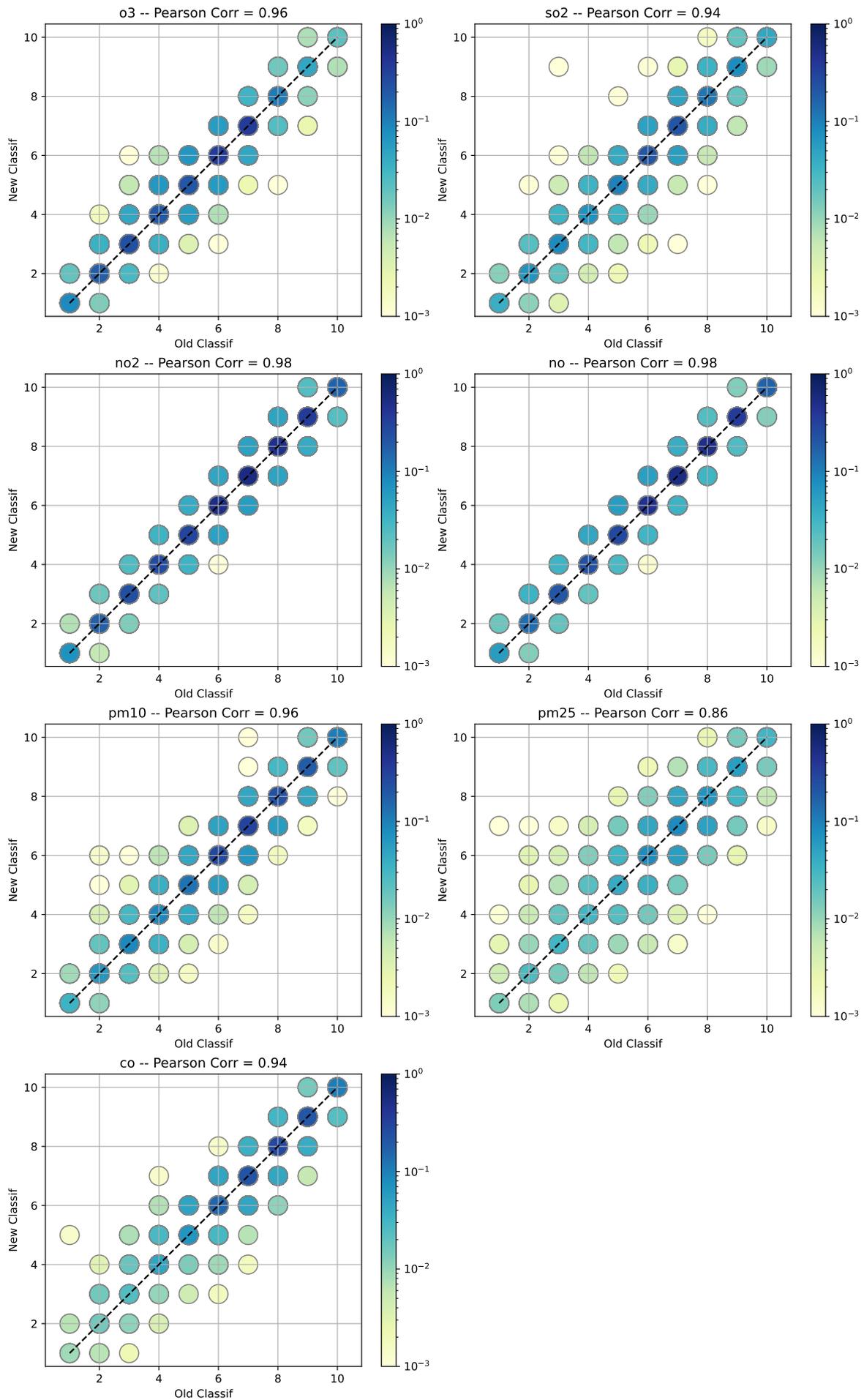
Figure 6 – Validation croisée : nombre et pourcentage (en couleur) dans chaque classe pour chaque type de station. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.



**Figure 7** – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.



**Figure 8** – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.



**Figure 9** – Scatter Plot des classes obtenues avec l'ancienne et la nouvelle classification. La couleur indique la fréquence d'occurrence.

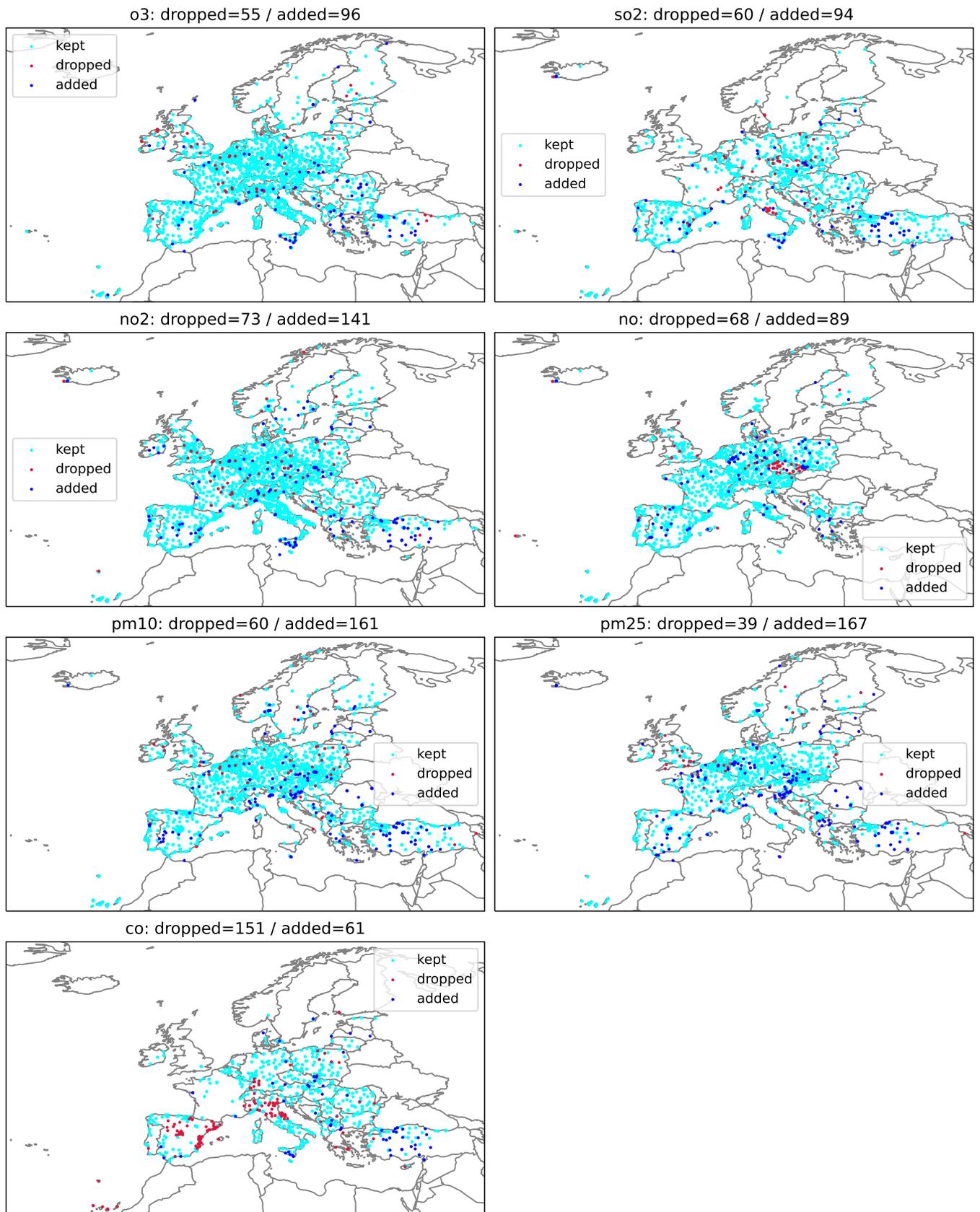
## 7 Évolution du jeu de stations classifiées

La figure 10 permet de suivre l'évolution du jeu de données classifiées. On notera l'apparition de nouvelles stations en Turquie, et la disparition de stations mesurant le CO de façon grossière (éliminées en amont) , principalement en Espagne et en Italie.

## 8 Conclusion

Cette version utilise le flux de l'EEA mis en place dans le cadre de CAMS. La période d'étude comprend 8 années, avec des données non validées pour 2022 et 2023.

- Le réseau de mesure s'est légèrement étoffé, malgré l'élimination des données de CO à trop faible résolution.
- Comme dans la version précédente, la qualité des séries temporelles est insuffisante en certaines régions d'Italie, ainsi qu'Allemagne, Tchéquie et Grande Bretagne : les valeurs absentes sont trop nombreuses au sein de chaque journée.
- La cohérence entre les métadonnées et la classification objective est globalement stable.



**Figure 10** – Stations qui disparaissent (rouge), ou qui apparaissent (bleu) dans la nouvelle version.