Annexe4 :

Produits d'observation développés à DSO/CEP

La fusion de données – dans le sens confronter et intégrer des informations d'origines multiples dans le but de fournir une information encore plus précise – s'est particulièrement développée ces dernières années à la Direction de Systèmes d'Observation de Météo-France, pour notamment répondre à une demande croissante en information spatialisée. Ce type d'information a un intérêt évident dans le domaine de la prévision ou de la climatologie, en renseignant sur tout un domaine et non plus seulement aux endroits des stations de mesure. Cette évolution est intervenue dans un contexte d'optimisation des réseaux d'observation, avec le souci principal de valoriser les données existantes plutôt que de multiplier les sites de mesures *in situ*. Dans un tel contexte, les utilisateurs – qui étaient déjà à l'initiative de cette évolution – sont désormais incités à favoriser les données spatialisées produites à partir de différentes sources d'information plutôt que des observations ponctuelles.

Les données utilisées dans les différentes productions sont d'origines diverses et de natures différentes. On peut distinguer différentes sources d'information : de l'observation *in situ* (humaine ou automatique) aux données télédétectées (satellites, radars, capteurs foudre) en passant par des champs issus de la modélisation. L'information apportée peut être ponctuelle ou spatialisée, brute (sortie directe d'un capteur) ou élaborée (traitement d'image, assimilation de données). Ainsi on peut être amené à utiliser une température mesurée par un capteur ou la localisation d'un impact de foudre, mais aussi une image de classification nuageuse par satellite, une estimation des précipitations par radar ou un champ analysé de la température à 2 m. Enfin, les données peuvent être disponibles à des fréquences différentes (5 min, 15 min, 1 h, …) et l'information apportée peut être de type instantané ou bien intégrée sur toute une période (cumuls par exemple). La production étant clairement orientée vers l'observation, il faut noter que l'utilisation de données provenant de la modélisation est réduite au strict minimum, c'est à dire à l'utilisation de champs analysés résultant d'une assimilation de données d'observation.

En permettant généralement de fournir une information spatialisée, la fusion de données a ouvert de nouvelles perspectives dans différents domaines de la météorologie, de la prévision immédiate à la climatologie, en passant par le contrôle de modèle. Les différents produits développés répondent ainsi à des besoins parfois nouveaux ou dont la réalisation n'avait jusque-là pas été rendue possible. Au premier rang des utilisateurs exprimant des besoins, on trouve naturellement les prévisionnistes, qui souhaitent disposer d'informations supplémentaires pour les assister dans le court terme, à mi-chemin entre l'observation et la prévision. Ce sont les premiers utilisateurs potentiels des produits une fois qu'ils sont disponibles. Les produits peuvent ensuite – sous réserve qu'ils soient archivés – entrer dans le champ d'autres applications, parmi lesquelles les études ou l'utilisation climatologique. La plupart des produits développés le sont dans cette optique : tout d'abord fournir une aide supplémentaire aux prévisionnistes, et ensuite constituer une archive pour les études en temps différé.

Suite aux différents besoins exprimés, les produits suivants ont été développés :

- mention de l'activité convective dans les METAR automatiques (MACMA),
- analyse horaire des précipitations (ANTILOPE),
- analyse des chutes de neige en cours et de la tenue de la neige au sol (VISON),
- analyse du risque de brume et brouillard (CARIBOU),
- analyse de la visibilité (CERVUS).

Selon le besoin exprimé ou la capacité à y répondre, l'information apportée peut l'être soit sous forme quantitative (ex : quantité de précipitations, probabilité de visibilité inférieure à un seuil), soit sous forme qualitative (ex : risque faible, moyen ou fort de brume et brouillard).

Par ailleurs, une réanalyse des lames d'eau horaires sur la période 1997-2006 est en cours d'élaboration, tout comme une réflexion sur l'élaboration d'un produit relatif à la couverture nuageuse. La réanalyse des lames d'eau peut être considérée comme un prolongement d'ANTILOPE d'un point de vue de l'expression du besoin ; cependant cette production a pour objectif la constitution d'une archive de données, conçue pour fonctionner uniquement à partir de données archivées et non en temps réel, ce qui la différencie des autres produits développés jusque-là.
De manière générale, et pour répondre aux besoins des premiers utilisateurs que sont les prévisionnistes, la production doit se faire en temps réel. Il en découle un certain nombre de contraintes fortes sur le type de production mis en place. Ainsi, le délai de mise à disposition du résultat doit généralement être assez court, lequel peut également être contraint par la fréquence de production souhaitée. À l'inverse, la fréquence de production proposée aux utilisateurs peut être définie à partir du temps nécessaire à la production, lorsque la marge de manœuvre sur celui-ci est faible. En effet, les informations en entrée du produit ne sont pas forcément disponibles en nombre et à fréquence très élevés. Un nombre minimal de données en entrée peut alors être jugé nécessaire, lequel devient un facteur limitant la réduction du délai de mise à disposition et l'augmentation de la fréquence de production. Enfin, une autre contrainte à laquelle la production doit s'adapter vient de la qualité des données d'entrée elles-mêmes. Si ces données ont pu bénéficier d'une expertise humaine ou passer différents contrôles lorsque l'on se place dans le cadre d'une production en temps différé (réanalyse, reconstitution d'une archive), ce n'est pas le cas lors d'une production en temps réel où les données peuvent parfois être douteuses. Dans ce cas, l'implémentation de contrôles peut être requise pour un bon fonctionnement du produit.

Les méthodes employées pour réaliser la fusion de données proprement dite diffèrent d'un produit à l'autre, en fonction de la nature même du produit et des données d'entrée disponibles. Ainsi les méthodes retenues pour renseigner l'activité convective dans un METAR automatique (information locale, dans un rayon de 30 km) ou pour estimer une lame d'eau spatialisée vont être de nature différente. Ensuite, la complexité de la méthode va également être influencée par le degré du lien entre les données d'entrée disponibles et la grandeur physique recherchée en sortie, ainsi que par la densité de ces données d'entrée lorsqu'elles sont ponctuelles.

Ainsi le produit MACMA utilise naturellement l'imagerie radar et les impacts de foudre détectés, lesquels sont directement liés au résultat voulu. Et alors un simple croisement des deux types d'information s'avère suffisant. On retrouve une certaine simplicité dans le produit CARIBOU, où les liens entre les données d'entrée et le phénomène étudié (la brume ou le brouillard) sont supposés connus par l'intermédiaire de critères simples favorables ou défavorables à l'occurrence du phénomène, et où les données d'entrée sont déjà spatialisées. Il suffit alors juste de vérifier ces critères de manière séquentielle en tout point du domaine considéré.

La complexité intervient surtout lorsque l'on veut mêler informations ponctuelles et spatialisées. Par exemple, pour estimer une lame d'eau, on pense tout de suite à utiliser images radar et données pluviométriques au sol, les deux fournissant déjà des quantités de précipitations. La difficulté réside ici dans la combinaison de ces deux informations. Alors deux types d'approche peuvent être envisagés : soit on considère que les deux sources doivent être utilisées de manière plutôt disjointes (comme dans ANTILOPE qui sépare pluies stratiformes et convectives, en utilisant respectivement les données pluviométriques et les données radar), soit on cherche directement à les mélanger (typiquement, en ajustant les données radar à l'aide des pluviomètres), ce qui est l'option choisie dans le cadre de la réanalyse des lames d'eau. Dans les deux cas, il faut recourir à une méthode de spatialisation, en l'occurrence ici le krigeage (ordinaire ou à dérive externe), méthode géostatistique éprouvée et largement utilisée dans le domaine de l'estimation des précipitations.

Une difficulté supplémentaire apparaît lorsque que la donnée spatialisée et les données ponctuelles que l'on veut fusionner ne mesurent par la même grandeur ou ne sont pas du même type. C'est le cas

dans le produit VISON qui va chercher par exemple à combiner un champ continu de température et des informations ponctuelles de type binaire : il neige ou non, il y a de la neige au sol ou non. Il est décidé dans ce cas de transformer ces dernières afin de les assimiler à des températures, en leur attribuant une valeur qui rend possible ou pas le phénomène observé. Ceci permet ensuite de réaliser un krigeage avec dérive externe. Dans le cas précis d'un champ de température, dont l'existence d'un lien avec l'altitude est connue, il apparaît naturel de tenir compte de cette liaison dans la phase de spatialisation. Ainsi il est fait recours dans VISON à du krigeage avec dérive externe pour obtenir des champs de températures (à 2 m ou au sol) à partir d'observations ponctuelles et en prenant des ébauches de champs de température et l'altitude comme variables spatialisées externes.
Ce recours à des données de type géophysique (comme le relief) est particulièrement mis en avant dans le produit CERVUS et se justifie notamment par le faible nombre d'observations du phénomène étudié – ici la visibilité – pour lequel on ne dispose par ailleurs pas d'ébauche spatialisée (comme une image radar pour la pluie ou une analyse modèle pour la température). Il paraît inconcevable de fournir une estimation de qualité en spatialisant directement les quelques observations de visibilité disponibles, que ce soit par krigeage ou interpolation avec l'inverse de la distance. Mais si l'on parvient à mettre en lumière une relation statistique entre ces observations et différentes données spatialisées – de type météorologique ou géophysique – alors on peut envisager une nouvelle méthode de spatialisation, en appliquant cette relation en tout point du domaine. Le champ obtenu peut alors soit être conservé tel quel, soit servir d'ébauche. C'est ce qui est fait dans CERVUS où on a recours à la régression linéaire multiple avec la visibilité comme prédictand et différents prédicteurs météorologiques (classification nuageuse, précipitations, analyses de pression, température, humidité et vent) ou géophysiques (relief, nature du sol, élévation solaire). Le champ de visibilité obtenu sert d'ébauche qui est ensuite corrigée à l'aide des visibilités observées.

Le dernier aspect de la production des différents produits – et pas le moindre – est leur validation. Celle-ci doit se faire avant la mise en production opérationnelle, autant que possible avec les utilisateurs finaux. Des protocoles de validation sont mis en place pour des évaluations subjectives par les premiers utilisateurs (prévisionnistes généralement), lesquels accèdent aux données par l'intermédiaire de maquettes web créées pour l'occasion. Les retours obtenus sur une période couvrant quelques mois sont analysés et peuvent permettre quelques premières adaptations du produit. En parallèle, un protocole de validation objective est mis en place. Cette validation plus systématique faisant intervenir différents scores statistiques permet de mieux qualifier le produit et de faire apparaître des limites au produit, lesquelles – si elles ne peuvent pas être corrigées – sont portées à la connaissance des utilisateurs dans un document spécifique.

Une difficulté particulière réside notamment dans la validation des produits sur certaines zones (mer, ou montagne) renseignées *a priori* par le produit mais où on ne dispose pas forcément d'éléments pour valider. Il est alors décidé soit de masquer les données sur certaines zones, soit d'informer l'utilisateur que la validité du produit n'y est pas assurée. Dans la pratique actuelle, c'est plutôt l'option du masquage de données qui est privilégiée, mais les utilisateurs ont de plus en plus tendance à vouloir de l'information partout, quitte à ce qu'elle soit de qualité incertaine. Cette tendance conforte l'idée que la donnée doit être accompagnée d'une information relative à sa qualité, que l'on peut appeler qualification des données. Même si ce n'est pas quelque chose de généralisé actuellement, certains produits sont déjà ou vont être accompagnés d'une information sur la qualité des données fournies. ANTILOPE fournit notamment un code qualité combinant code qualité radar et variance d'erreur de krigeage, et la réanalyse des lames d'eau sera accompagnée d'une incertitude sur l'estimation. D'autres produits comme CERVUS sont fournis de manière probabiliste et intègrent donc déjà en partie cette notion d'incertitude.

De manière générale, les produits de fusion de données présentent l'avantage d'être assez robustes dans le sens où ils sont censés fournir une information même en l'absence de certaines données

d'entrée, du fait de l'utilisation d'un maximum de sources de données. Le danger serait justement qu'un produit soit trop dépendant d'une de ses données d'entrée au point où il ne pourrait pas fonctionner si elle venait à ne pas être disponible, que ce soit de manière ponctuelle ou définitive.