Validation et étude de quelques propriétés de systèmes de prévision météorologique ensemblistes

Frédéric Atger

Université Toulouse III - Paul Sabatier

THESE

pour obtenir le grade de

Docteur de l'université Toulouse III

Discipline : météorologie

Présentée et soutenue par

Frédéric Atger

le 14 avril 2003

Validation et étude de quelques propriétés de systèmes de prévision météorologique ensemblistes

Directeur de thèse : Olivier Talagrand

En présence de :

Robert Rosset	Président du jury
Eric Parent	Rapporteur
Zoltan Toth	Rapporteur
Robert Vautard	Rapporteur
Philippe Bougeault	Examinateur
Michel Déqué	Examinateur
Bernard Strauss	Examinateur

Remerciements

Mes remerciements vont d'abord à Olivier Talagrand. Il a su se montrer attentif, patient et enthousiaste, malgré l'éloignement géographique, un emploi du temps chargé et le rythme très irrégulier de mon travail. Ce fut un honneur de travailler sous sa bienveillante houlette.

Michel Déqué a été un « parrain » consciencieux. Le hasard a voulu qu'il dirigeât, il y a 17 ans, mon stage de fin d'études d'ingénieur de l'ENM. Ce n'était donc qu'un juste retour des choses qu'il suive mon travail durant ces trois dernières années.

Merci à Philippe Bougeault pour sa participation au jury, dont Robert Rosset a eu la gentillesse d'accepter la présidence.

Je suis sincèrement reconnaissant à Eric Parent, Robert Vautard et Zoltan Toth d'avoir accepté la charge de relecture de ce mémoire.

Maurice Merlet et Bruno Lacroix m'ont permis de reprendre mes travaux de recherche tout en continuant d'occuper un poste à la Direction de la Prévision de Météo-France, qu'il m'aurait été pénible de quitter. Mes collègues de la division COMPAS, surtout ceux de l'équipe *adaptation statistique*, ont dû s'accommoder de ma disponibilité réduite. Qu'ils en soient tous ici remerciés.

Gwenaëlle Hello et Jean-Noël Thépaut ont contribué, peut-être sans le savoir, à ce que je me décide à m'inscrire à l'Université pour entreprendre cette thèse.

Je remercie enfin Bernard Strauss, également membre du jury, sans qui cette thèse n'aurait jamais vu le jour. En m'accueillant dans son équipe au CEPMMT il a permis au prévisionniste que j'étais de se frotter au monde de la recherche, jusqu'à y prendre goût. Il m'a encouragé dans mes travaux bien avant qu'il soit question d'en faire une thèse, et il fut l'initiateur des premiers « ensembles du pauvre » à partir desquels sont bâtis les deux premiers chapitres de cette thèse.

Table des matières

Introduction 9

Historique des travaux réalisés 29

Chapitre 1 Performance de prévisions probabilistes basées sur des distributions construites à partir de différents modèles et systèmes de prévision d'ensemble 37

1. Introduction 37

2. Méthode et définitions 40

3. Principaux résultats 41

3.1. La performance de l'ensemble du CEPMMT 41

3.2. Contributions de la moyenne et de la dispersion de l'ensemble 42

3.3. Impact du nombre de membres de l'ensemble 42

3.4. Comparaison entre les ensembles du CEPMMT et du NCEP 43

3.5. Ensemble multi-modèles 43

4. Remarques conclusives 44

Bibliographie 45

Article (publié) : *The skill of ensemble prediction systems* (1999) 47

Chapitre 2 Valeur économique de prévisions probabilistes de forts cumuls de précipitation sur la France 61

1. Introduction 61

2. Méthode 67

3. Principaux résultats 68

3.1. Comparaison ensemble - contrôle 68

3.2. Comparaison ensemble - ensemble 69

4. Discussion 70

5. Remarques complémentaires 71

Annexe 72

Bibliographie 74

Article (publié) : *Verification of intense precipitation forecasts from single models and ensemble prediction systems* (2001) 77

Chapitre 3 Variabilité spatiale et temporelle de la fiabilité de prévisions probabilistes issues d'un ensemble Conséquences pour la calibration 95

1. Introduction 95

2. Méthode 100

3. Principaux résultats 102

3.1. Variabilité inter-annuelle de la fiabilité 102

3.2. Variabilité spatiale de la fiabilité 102

3.3. Variabilité inter-annuelle de la fiabilité locale 103

4. Discussion 103

4.1. L'estimation de la fiabilité 103

4.2. Significativité des résultats 104

4.3. Variations du biais et de la fiabilité 104

4.4. Conséquences pour la calibration 105

Annexe 106

Bibliographie 108

Article (accepté) : *Spatial and interannual variability of the reliability of ensemble based probabilistic forecasts - Consequences for calibration* 111

Chapitre 4 Estimation de la fiabilité de prévisions probabilistes à partir d'échantillons de taille réduite 141

1. Introduction 141

2. Données et définitions 145

3. Impact de la catégorisation 145

4. Utilisation d'un test statistique 146

5. Estimation par ajustement de la courbe ROC 146

6. Conséquences pour la calibration 147

7. Remarques conclusives 148

Annexe 149

Bibliographie 151

Article (soumis) : *Estimation of the expected reliability of ensemble based probabilistic forecasts* 153

Conclusion et perspectives 181

Introduction

L'incertitude des prévisions météorologiques préoccupe depuis toujours ceux qui entreprennent de prévoir, par des moyens scientifiques, le temps qu'il va faire dans les heures ou les jours à venir. Prévoir, c'est aussi vérifier en permanence qu'on ne se trompe pas trop souvent. C'est donc aussi se poser la question, au moment même où la prévision s'élabore, de la qualité de la prévision. Les prévisionnistes savent depuis toujours reconnaître des situations « difficiles », pour lesquelles l'évolution de l'atmosphère est, davantage qu'à l'accoutumée, entâchée d'incertitude. Au contraire, des situations de grande stabilité ont été identifiées comme permettant d'étendre les échéances de prévision au-delà de ce qu'il est raisonnable de pratiquer en temps normal (Atger, 2000). De leur côté, les modélisateurs ont constaté que l'évolution prévue par les modèles numériques est plus ou moins sensible, suivant les cas, à de petites modifications apportées à l'état initial (Rabier et al., 1996).

En vérité, nous ignorons jusqu'à quel point l'évolution atmosphérique est sensible à ses conditions initiales, faute d'un dispositif expérimental à l'échelle planétaire. En revanche, nous savons que les petites imperfections de l'état initial (mais aussi, dans une moindre mesure, les approximations apportées à la modélisation de certains processus physiques) dégradent plus ou moins rapidement les prévisions numériques. L'atmosphère, ou plus précisément son avatar modélisé, présente les caractéristiques d'un système chaotique, telles que Poincaré (1914) les a décrites il y a presque un siècle, avant que Lorenz (1963) ne les remette au goût du jour grâce aux ressources offertes par les moyens modernes de calcul numérique. Si l'évolution atmosphérique est en grande partie imprévisible au-delà de quelques jours d'échéance, ce n'est pas tant que les modèles utilisés soient imparfaits, mais surtout en raison de notre incapacité à connaître et décrire parfaitement les conditions initiales.

Faut-il pour autant parler de prévisibilité intrinsèque de l'atmosphère ? Probablement pas, car c'est à travers l'intégration par un modèle de prévision que se matérialise la croissance des erreurs de l'état initial. Cette croissance est évidemment affectée par les particularités du modèle utilisé. En outre, la qualité de la collecte et du traitement des observations, opérations qui sont spécifiques à chaque modèle (ou plus précisément à chaque système de prévision construit autour d'un modèle), ont un impact sur l'incertitude relative aux conditions initiales. On ne peut donc parler de prévisibilité que relativement à un système de prévision numérique. Il est raisonnable de chercher à connaître la sensibilité d'une prévision numérique à différentes sources d'erreur, et en premier lieu à l'erreur de l'état initial, pour estimer au cas par cas le niveau d'incertitude. Cette opération est généralement assimilée à une prévision de la qualité de la prévision (Palmer & Tibaldi, 1988). Une telle prévision ne saurait cependant être déterministe, car le niveau de l'incertitude ne conditionne évidemment pas le niveau de l'erreur[1]. Elle permet en revanche d'apprécier le niveau de confiance qu'il faut accorder à la prévision, ce qui est le moins qu'on puisse attendre d'une prévision dont le caractère radicalement incertain ne fait aucun doute.

§§§

La nécessité d'assortir une prévision d'une estimation de son niveau d'incertitude est justifiée au plan épistémologique par Popper (1982) dans des considérations relatives aux exigences de l'approche scientifique. La prévision est considérée comme le résultat de l'application d'une théorie au monde réel, application visant à déduire l'état futur à partir de l'état présent. Ce n'est pas l'inadéquation de la théorie (qui reste aux yeux de Popper une pure conjecture) qui impose de compléter la prévision par une information relative à son niveau d'incertitude,

[1]L'expression « forecast of the forecast skill » n'est cependant pas synonyme de « forecast of the forecast error ». Le « skill » n'est pas non plus la qualité, c'est un terme plus précis qui peut être traduit par « aptitude » et qui pourrait, dans l'expression ci-dessus, prendre le sens de « capacité à prévoir », c'est à dire précisément ce qu'indique le niveau d'incertitude. Il persiste cependant une ambiguïté sur le sens précis que les auteurs accordent au terme de « skill », qui peut dans un même article représenter l'aptitude d'un système de prévision, mais

mais l'impossibilité dans laquelle nous nous trouvons d'accéder à une connaissance exhaustive du monde réel. Plus précisément, Popper affirme qu'un scientifique doit être capable d'indiquer le degré de précision qu'il faut atteindre dans la connaissance de l'état initial pour obtenir une prévision dont le degré de précision est imposé a priori.

De façon plus pragmatique, le principal défaut d'une prévision déterministe est qu'elle ne permet pas d'apprécier le risque que les indications fournies soient erronées. C'est pourtant le niveau de ce risque qui conditionne la manière dont un utilisateur va tirer profit de la prévision. Dans le cas simple de la prévision d'occurrence d'un phénomène météorologique, chaque utilisateur est caractérisé par un certain niveau de tolérance à deux types d'erreurs de prévision : la non détection, c'est à dire l'absence d'avertissement précédant l'occurrence du phénomène ; la fausse alarme, c'est à dire l'avertissement injustifié qui conduit à prendre des mesures qui s'avèrent inutiles. Par exemple, la tolérance à la fausse alarme est généralement plus élevée pour la prévision d'une violente tempête que pour l'annonce d'un temps faiblement pluvieux : on renonce plus facilement à une sortie en mer, pour ne pas risquer d'y laisser la vie, qu'on n'abandonne un projet de promenade en montagne pour éviter d'être mouillé. Chaque utilisateur, suivant sa sensibilité (et ses aptitudes à la navigation maritime et à la marche en montagne) aura cependant une stratégie spécifique dans chacune des ces deux situations.

En ne fournissant aucune information relative à l'incertitude de la prévision, on empêche l'utilisateur de se positionner par rapport au risque de fausse alarme et de non détection. Par ailleurs, la réalisation d'une prévision déterministe impose au prévisionniste de faire des hypothèses, rarement explicites, quant à la sensibilité des utilisateurs aux deux types d'erreur mentionnés ci-dessus. Lorsqu'il existe un risque, même limité, de phénomène grave ou dangereux, cette

---

aussi la qualité d'une prévision produite par ce système (par exemple : Tennekes, 1991).

attitude conduit généralement à une « sur-protection » (Murphy, 1991), certainement nécessaire sur le plan social, mais qui s'avère dommageable à la fois pour les utilisateurs, qui prennent individuellement une décision sous-optimale, et pour le prévisionniste dont la crédibilité résiste mal à des fausses alarmes répétées. On peut en conclure qu'une prévision déterministe est tout simplement tronquée de sa composante probabiliste (Murphy, 1993), si on admet que ce dernier terme n'implique pas nécessairement l'introduction de probabilités d'occurrence de certains événements, mais seulement d'indications quant à l'incertitude de la prévision. Certes la composante probabiliste est-elle toujours présente implicitement, si on considére que l'utilisateur sait d'expérience le niveau d'incertitude des prévisions qui lui sont fournies. Formuler explicitement un risque permet cependant d'indiquer comment varie jour après jour le niveau d'incertitude, et c'est bien-sûr cette information qui possède la plus grande valeur pour l'utilisateur de la prévision.

Ce n'est donc pas seulement une formulation probabiliste qui est recherchée, formulation qui pourrait consister à transformer une prévision déterministe en une probabilité, à partir de statistiques portant sur l'erreur de prévision constatée pendant une période suffisamment longue, mais bien une estimation de l'incertitude qui soit conditionnelle à la prévision. Encore faut-il disposer de moyens pour effectuer cette estimation.

§§§

L'estimation du niveau d'incertitude d'une prévision peut être purement subjective. Il n'y a pas en Europe de tradition établie, chez les météorologistes, d'élaboration de prévisions probabilistes. Aux Etats-Unis, en revanche, la formulation probabiliste des prévisions est généralisée, au moins pour ce qui concerne l'occurrence de précipitations (Hamill & Wilks, 1994). Cette formulation repose le plus souvent sur une estimation subjective des risques par le prévisionniste. Il semble que cette estimation fasse partie intégrante du processus hypothético-déductif par lequel un prévisionniste est conduit à

formuler une prévision (Sanders, 1963). Les opinions du prévisionniste ne sont pas déterministes, si on considère sous ce terme les représentations mentales qui sont à l'oeuvre au cours de son travail d'expertise (Murphy, 1991). Disposant de sources d'information multiples et d'origines diverses, le prévisionniste envisage différentes options, parfois contradictoires, avant de conclure sous une forme plus ou moins déterministe. On peut d'ailleurs considérer qu'il s'agit d'une caractéristique du travail d'expert que de se forger des opinions probabilistes dans un premier temps, c'est à dire d'estimer des risques, et de rendre dans un second temps un jugement déterministe. Le caractère déterministe du jugement final n'est cependant qu'une contrainte sociale, extérieure au processus d'élaboration proprement-dit.

§§§

L'estimation du niveau d'incertitude de la prévision peut faire appel à des procédures statistiques. Si le niveau d'incertitude varie chaque jour, en fonction des conditions initiales et de la capacité du modèle à les faire évoluer, on doit pouvoir trouver des prédicteurs parmi les variables décrivant ces conditions initiales et leur évolution prévue, et modéliser la relation statistique qui existe entre ces prédicteurs et le niveau d'erreur de la prévision. Une telle approche a été suivie au CEPMMT à la fin des années 1980, sans grand succès (Molteni & Palmer, 1988). Un modèle statistique linéaire était ajusté à partir de données historiques et permettait de fournir chaque jour une estimation du niveau d'incertitude en quelques classes. Pour avoir vécu cette expérience « sur le terrain » (j'étais prévisionniste à cette époque) je considère que l'échec de cette tentative fut davantage la conséquence d'un malentendu sur ce que la procédure statistique était supposée fournir (l'incertitude n'est pas l'erreur), que sur la pertinence des estimations fournies quotidiennement.

Des prévisions probabilistes, obtenues par traitement statistique des variables atmosphériques prévues par l'intégration unique d'un modèle, sont d'ailleurs produites en routine dans de nombreux pays (par exemple : Carter et al., 1989),

et sont utilisées quotidiennement par les prévisionnistes. Le produit fourni n'est pas estampillé « prévision de la qualité de la prévision », mais il s'agit bien d'une prévision du niveau d'incertitude par des moyens statistiques, qui semble donner toute satisfaction aux utilisateurs (c'est le cas à Météo-France pour des probabilités de brouillard et de forte rafale, en cours de validation).

§§§

Les systèmes de prévision d'ensemble représentent une troisième approche pour l'estimation du niveau d'incertitude des prévisions. La modélisation explicite de la distribution de probabilité (pdf)[2] de la prévision à partir de la pdf de l'état initial (à l'aide de l'équation de Liouville ou de l'équation de Fokker-Planck), est théoriquement possible mais pratiquement exclue pour ce qui concerne les modèles numériques utilisés pour la prévision météorologique, du fait de la haute dimensionnalité de l'espace considéré pour modéliser l'atmosphère (Ehrendorfer, 1994). Des techniques d'échantillonnage de type « Monte-Carlo » ont été proposées pour pallier cette limitation. Elles consistent en principe à intégrer le modèle de prévision à partir d'un grand nombre d'états initiaux, obtenus à partir de l'état initial de référence en le modifiant de façon aléatoire dans les limites de l'incertitude relative à son estimation. Ces états initiaux *perturbés* constituent un échantillonnage de la pdf de l'état initial, et leur intégration est supposée échantillonner la pdf de l'état final. En pratique, le coût de multiples intégrations du modèle numérique est rapidement prohibitif. Le nombre d'états initiaux perturbés qui peuvent être pris en considération est limité à quelques dizaines dans le meilleur des cas, ce qui rend impossible un échantillonnage réaliste de la pdf de l'état initial, la plupart des perturbations générées de manière aléatoire évoluant en dehors de l'attracteur du système (Hollingsworth, 1980).

Ce sont finalement des techniques d'échantillonnage sélectif qui ont permis de produire des prévisions d'ensemble opérationnelles. Ces techniques reposent sur

[2] Probability density function.

l'identification d'un petit nombre de perturbations qui sont supposées représenter les éléments les plus importants pour l'évolution de l'incertitude sur l'état de l'écoulement. Deux méthodes concurrentes ont vu le jour pour réaliser cet objectif. La première repose sur le calcul des « vecteurs singuliers » (Mureau et al., 1993) et permet d'identifier les structures dynamiques qui *seront* les plus instables dans l'avenir proche tel que le modèle le prévoit, et qui sont donc susceptibles d'indiquer des zones de sensibilité de l'état initial. La seconde méthode est basée sur la « culture[3] » de perturbations ('bred modes') au sein du processus d'analyse de l'état initial (Toth & Kalnay, 1993) et vise à identifier les structures dynamiques qui *ont été* les plus instables dans le passé et qui sont donc susceptibles d'être représentées de façon inexacte dans l'état initial. Ces deux méthodes sont utilisées respectivement par les deux systèmes de prévision d'ensemble concurrents qui sont opérationnels depuis 1993 : en Europe, au CEPMMT[4] (Palmer et al., 1993) et aux Etats-Unis au NCEP[5] (Tracton & Kalnay, 1993). Après des débuts difficiles, du fait de la résolution dégradée du modèle utilisé et du faible nombre de membres, conditions rendues toutes deux nécessaires par la limitation des capacités de calcul, ces deux systèmes sont devenus des outils importants, sinon essentiels, pour la prévision météorologique en Europe et aux Etats-Unis. D'autres systèmes de prévision d'ensemble ont vu le jour, prenant en compte non seulement l'incertitude relative à l'état initial mais aussi parfois les erreurs probables de modélisation (Houtekamer et al., 1996 ; Kobayashi et al., 1996 ; Hersbach et al., 2000 ; Hou et al., 2001 ; Mendonça & Bonatti, 2002 ; Naughton et al., 2002 ; Nicolau, 2002 ; Tian et al., 2002).

§§§

[3] « Breeding » est généralement traduit par « élevage », mais le terme de « culture » évoque mieux le mécanisme par lequel une perturbation aléatoire est d'abord « semée » puis subit plusieurs cycles de croissance jusqu'à atteindre la maturité.

[4] Centre Européen pour les Prévisions Météorologiques à Moyen Terme, Reading, Royaume-Uni.

[5] National Centers for Environmental Prediction, USA.

En marge du développement de ces systèmes opérationnels ou expérimentaux, l'évaluation de la qualité des ensembles est devenue un champ de recherche très actif qui a donné lieu à de nombreuses publications au cours des dernières années (Buizza, 1997 ; Du et al., 1997 ; Hamill & Colucci, 1997 ; Petroliagis et al., 1997 ; Talagrand et al., 1997 ; Buizza et al., 1998 ; Buizza & Palmer, 1998 ; Eckel & Walters, 1998 ; Hamill & Colucci, 1998 ; Whitaker & Loughe, 1998 ; Buizza et al., 1999 ; Hamill, 1999 ; Molteni & Buizza, 1999 ; Wilson et al., 1999 ; Chessa & Lalaurette, 2000 ; Hersbach, 2000 ; Richardson, 2000 ; Toth et al., 2000 ; Ziehmann, 2000 ; Buizza, 2001 ; Hamill, 2001 ; Hou et al., 2001, Mullen & Buizza, 2001 ; Richardson, 2001 ; Zhu et al., 2001 ; Buizza & Chessa, 2002 ; Mullen & Buizza, 2002 ; Mylne, 2002 ; Szunyogh & Toth, 2002). Deux facteurs ont contribué à susciter des travaux de validation des systèmes de prévision d'ensemble : d'une part la concurrence entre les méthodes de génération des perturbations de l'état initial, dominée par un débat de fond sur les avantages respectifs des 'vecteurs singuliers' et des 'bred modes' ; d'autre part le constat que les prévisions d'ensemble coûtent cher et sont relativement peu utilisées en routine, même aujourd'hui, pour élaborer les prévisions finales qui sont fournies à l'immense majorité des utilisateurs (Atger, 2001). L'objectif de ces études est donc à la fois de valider l'approche scientifique suivie pour construire un ensemble, mais aussi de mettre en évidence la qualité des prévisions et leur utilité.

L'évaluation de la qualité des ensembles peut passer par des investigations plus ou moins approfondies portant sur les caractéristiques des perturbations apportées à l'état initial (amplitude, répartition spatiale, études de cas) ainsi que sur leur taux de croissance (Molteni et al., 1996). L'analyse en composantes principales est parfois utilisée pour mettre en évidence la capacité des perturbations à explorer différentes régions de l'attracteur (Molteni & Buizza, 1999). La plupart des études portent cependant sur l'évaluation de la qualité des prévisions. Les critères de qualité sont multiples, suivant qu'on privilégie l'un ou l'autre des objectifs assignés à la prévision d'ensemble. Un de ces objectifs est la

quantification de l'incertitude de prévision, qui donne lieu à des évaluations de la relation entre la dispersion de l'ensemble et la performance d'une prévision déterministe issue ou non de cet ensemble[6] (Buizza & Palmer, 1998), évaluations qui concluent généralement au caractère sous-dispersif des ensembles opérationnels. Cette tendance à sous-estimer la dispersion est souvent mise en évidence par le biais d'histogrammes de rang[7], qui indiquent, entre autres diagnostics statistiques, l'aptitude de l'ensemble à « englober » la vérification (Talagrand et al., 1997 ; Hamill, 2001). Un autre objectif des systèmes de prévision d'ensemble est l'identification de scénarios météorologiques alternatifs, dont la probabilité *a posteriori* doit évidemment être en rapport avec le nombre de membres de l'ensemble qui les représentent (Molteni et al., 1996). Le niveau moyen de performance de la moyenne d'ensemble est également beaucoup utilisé comme critère de qualité (Buizza & Palmer, 1998). L'intérêt d'utiliser la moyenne d'ensemble a été mentionné bien avant qu'on dispose d'ensembles opérationnels (Leith, 1974). L'opération de moyenne résulte en effet en un filtrage des champs météorologiques, dont on peut penser qu'il est optimal puisque dépendant de la dispersion des membres, et qui permet de ne conserver que les structures météorologiques prévisibles, tandis que les phénomènes de petite échelle dont la prévision est incertaine disparaissent.

§§§

Le principal objectif assigné aux systèmes de prévision d'ensemble reste cependant la fourniture de prévisions probabilistes objectives. Un ensemble n'étant rien d'autre qu'un échantillonnage d'une distribution de probabilité, il n'est pas étonnant que la plupart des études portant sur la validation des ensembles s'appuient sur la vérification de prévisions probabilistes. C'est

[6] Cette relation est souvent appelée la 'spread-skill relationship', terme dont la traduction exacte est malaisée du fait de la signification ambiguë du mot 'skill' (ambiguïté discutée dans une précédente note). La force de cette relation est généralement évaluée par le niveau de correspondance entre la variance de l'ensemble et l'erreur quadratique moyenne de la moyenne de l'ensemble.

[7] 'Rank histogram', aussi appelé diagramme de Talagrand.

également le cas des travaux présentés dans cette thèse. Le sens commun veut qu'une prévision probabiliste ne soit jamais ni juste, ni fausse. Il est vrai que ce n'est qu'en accumulant un certain nombre de cas qu'on peut espérer conclure quant à la qualité d'un système de prévision probabiliste. Car la réalité reste unique face à une pdf échantillonnée par les membres de l'ensemble.

D'une façon très générale, la vérification consiste en l'examen de la distribution conjointe d'une prévision et d'une vérification (Murphy & Winkler, 1987). Dans le cas d'un ensemble, on cherche la correspondance entre une distribution d'évolutions météorologiques (échantillon supposé de la pdf) et une évolution de référence observée ou analysée. Ces deux entités n'étant pas de même nature, une catégorisation est nécessaire afin de vérifier que, si on regroupe tous les cas où une pdf particulière est prévue, la vérification est bien distribuée suivant cette pdf. Le problème étant multi-dimensionnel, et les échantillons de données généralement petits par rapport au nombre de degrés de liberté, on se contente le plus souvent de considérer certains aspects saillants de la pdf pour effectuer cette catégorisation. Un de ces aspects est la probabilité, conditionnelle à la pdf, qu'un phénomène météorologique donné se produise. Ainsi, la vérification de prévisions probabilistes issues d'un ensemble n'a pas seulement pour but d'évaluer la qualité de prévisions susceptibles d'être fournies à des utilisateurs, elle représente aussi une forme d'évaluation de la qualité de l'ensemble dont elles sont issues.

Le principal diagnostic utilisé pour évaluer la performance de prévisions probabilistes d'un événement est le score de Brier, largement utilisé dans cette thèse (Brier, 1950). Le score de Brier est l'erreur quadratique moyenne de la probabilité prévue, sous l'hypothèse que la probabilité parfaite vaut 1 quand l'événement considéré se produit, 0 quand l'événement ne se produit pas. Le score de Brier gagne à être décomposé en ses termes de fiabilité et de résolution (Murphy, 1973), le premier terme indiquant la correspondance entre probabilité prévue et fréquence observée (il doit pleuvoir dans 48% des cas où on a prévu une

probabilité de pluie de 48%) tandis que le second indique dans quelle mesure la fréquence observée varie quand la probabilité prévue change. Si la fiabilité est parfaite, la résolution indique la variabilité de la probabilité prévue, dont on comprend facilement que c'est en effet une condition pour qu'une prévision probabiliste présente quelque utilité. Le complément naturel du score de Brier est le diagramme de fiabilité qui représente graphiquement la relation entre probabilité prévue et fréquence observée, ainsi que la distribution des probabilités prévues.

La théorie du signal, utilisée à l'origine dans le champ de la médecine et de la psychologie expérimentale, a donné naissance à la courbe ROC (vraisemblablement pour Relative Operating Characteristics, encore que les auteurs, même anciens, ne s'accordent pas sur l'origine de cet acronyme) qui est largement utilisée pour l'évaluation de la qualité de prévisions probabilistes (Mason, 1982). L'information fournie par la courbe ROC présente l'avantage d'être bi-dimensionnelle, puisqu'elle indique les variations conjointes du taux de fausse alarme et du taux de détection de l'événement considéré. On a vu plus haut qu'un utilisateur se caractérise précisément par sa sensibilité à l'un ou à l'autre de ces indicateurs. La courbe ROC fournit ainsi un diagnostic qui peut facilement être personnalisé. C'est également le cas du diagnostic de valeur économique, qui peut être vu comme un prolongement de l'approche par la courbe ROC, faisant appel à un modèle simple de décision en situation d'incertitude (Richardson, 2000).

§§§

Les deux thèmes abordés dans la thèse sont d'une part la qualité relative de différents systèmes de prévision d'ensemble (chapitres 1 et 2), d'autre part l'estimation de la fiabilité dans le cas de prévisions probabilistes issues d'un ensemble (chapitres 3 et 4).

Dans les chapitres 1 et 2 est principalement étudiée la performance des prévisions probabilistes issues des ensembles opérationnels, par rapport à des prévisions probabilistes construites à partir de l'intégration unique d'un modèle de prévision, ou par rapport à des prévisions probabilistes issues d'un ensemble formé à partir de prévisions déterministes élaborées dans quelques centres de prévision opérationnelle. Le chapitre 1 présente une évaluation de prévisions d'un dépassement de seuil climatologique du géopotentiel à 500 hPa, à partir de la décomposition du score de Brier en ses termes de fiabilité et résolution. Dans le chapitre 2 on a recours à un diagnostic de valeur économique, appliqué à la vérification de prévisions de cumuls de précipitations. Le mode de vérification est original, en ce qu'il permet de tenir compte des incertitudes de localisation et d'intensité qui caractérisent les prévisions de fortes précipitations.

Le terme de fiabilité du score de Brier n'étant rien d'autre qu'une moyenne quadratique de biais catégoriques, son estimation est sensible aux problèmes d'échantillonage. Une accumulation importante de données est nécessaire pour l'évaluer, mais en même temps une estimation à partir d'un échantillon hétérogène présente peu d'intérêt. Le chapitre 3 porte sur la variabilité spatiale et temporelle de la fiabilité, et examine également les conséquences pour la calibration des prévisions probabilistes issues d'un ensemble. Dans le chapitre 4 on s'intéresse au problème de la catégorisation des probabilités prévues, catégorisation qui est à la base de la décomposition du score de Brier, et dont le choix généralement arbitraire a des conséquences importantes sur l'estimation du terme de fiabilité.

§§§

La présente introduction est suivie d'une synthèse de quelques pages, qui retrace l'historique de mes travaux de recherche et introduit brièvement les 4 chapitres

de la thèse, construits autour de 4 articles[8] en langue anglaise, rédigés entre fin 1997 et début 2003. Chaque chapitre s'ouvre par une introduction qui replace l'objet de l'article dans le contexte des différents thèmes abordés. Les points les plus importants, relatifs à la méthode et aux résultats obtenus, sont ensuite résumés et discutés. Il est fait systématiquement référence aux sections et aux figures de l'article, afin de faciliter la lecture de l'ensemble. L'article original est reproduit à la fin de chaque chapitre. La conclusion générale de la thèse reprend et discute certains des principaux résultats, et propose un certain nombre de perspectives.

§§§

Pour un meilleur confort de lecture on peut télécharger les articles originaux, ainsi que l'ensemble du manuscrit de thèse, sous forme de fichiers (au format *pdf*) depuis la page : http://www.multimania.com/fredatger/.

Bibliographie

Atger, F., 2000: La prévision du temps à moyenne échéance en France. *La Météorologie*, **8**, 30, 61-86.

Atger, F., 2001: Performance and usefulness of ensemble based probabilistic forecasts. 8th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 29-31).

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

[8] Deux de ces articles sont publiés (chapitres 1 et 2), le troisième est accepté (chapitre 3), le quatrième est soumis (chapitre 4).

Buizza, R., 1997: Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 99-119.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935-1960.

Buizza, R. and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503-2518.

Buizza, R., A. Hollingsworth, F. Lalaurette and A. Ghelli, 1999. Probabilistic predictions of precipitations using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168-189.

Buizza, R., 2001. Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev.*, **129**, 2329-2345.

Buizza, R. & P. Chessa, 2002. Prediction of the U.S. storm of 24-26 January 2000 with the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **130**, 1531-1551.

Carter, G.M., J.P. Dallavalle, and H.R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **4**, 401-412.

Chessa, P., and F. Lalaurette, 2000: Classification and verification of the ECMWF EPS perturbed forecasts using pre-defined regimes. *Wea. Forecasting*, **16**, 611-619.

Du, J., S.L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427-2459.

Eckel, F.A. and M.K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF Ensemble. *Wea. Forecasting*, **13**, 1132-1147.

Ehrendorfer, M., 1994. The Liouville equation and ist potential usefulness for the prediction of forecast skills. Part I: Theory. *Mon. Wea. Rev.*, **122**, 703-713.

Hamill, T. M & D. S. Wilks, 1994: A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting*, **10**, 620-631.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.

Hamill, T. M., and S. J. Colucci, 1998: Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Hamill, T. M, 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

Hamill, T. M, 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.*, **129**, 3, 550-560.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for Ensemble Prediction Systems. *Wea. Forecasting*, **15**, 5, 559-570.

Hersbach, H., R. Mureau & J. D. Opsteegh, 2000. A short-range to early-medium-range ensemble prediction system for the european area. *Mon. Wea. Rev.*, **128**, 3501-3519.

Hollingsworth, A., 1980. An experiment in Monte Carlo forecasting procedure. Workshop on stochastic dynamic forecasting, Reading, U.K., ECMWF (Proceedings).

Hou, D., E. Kalnay & K.K. Droegemeier, 2001. Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73-91.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie and H.L. Mitchell, 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Kobayashi, C., K. Yoshimatsu, S. Maeda, and K. Takano, 1996. Dynamical one-

month forecasting at JMA. Preprints, 11th Conf. on Numerical Weather Prediction, Norfolk, VA, Amer. Meteor. Soc., 13-14.

Leith, C.E., 1974. Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev.,* **102**, 409-418.

Lorenz, E.N., 1963. Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130-140.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

Mendonça, A.M. & et J.P. Bonatti, 2002. Ensemble global weather forecasting at CPTEC. WMO/CSB Technical Conference meeting, Cairns (Australia), December 2002 (Proceedings).

Molteni, F. & T. Palmer, 1988. An experimental scheme for the prediction of forecast skill at ECMWF. Predictability in the medium and extended range, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, 367-402).

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis, 1996. The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Molteni, F. & R. Buizza, 1999. Validation of the ECMWF ensemble prediction system using empirical orthogonal functions. *Mon. Wea. Rev.*, **127**, 2346-2358.

Mullen, S. L. and R. Buizza, 2001. Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-661.

Mullen, S. L. and R. Buizza, 2002. The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting*, **17**, 173-191.

Mureau, R., F. Molteni & T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.*, **119**, 299-323.

Murphy, A. H., 1973. A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

Murphy, A. H., 1991. Probabilities, odds, and forecasts of rare events. *Wea. Forecasting*, **6**, 302-307.

Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

Murphy, A. H. and R.L. Winkler, 1987. A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Mylne, K., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.*, in press.

Naughton, M., W. Bourke & G. Embery, 2002. Bureau of meteorology medium-range ensemble prediction system. WMO/CSB Technical Conference meeting, Cairns (Australia), December 2002 (Proceedings).

Nicolau, J., 2002. Short-range ensemble forecasting. WMO/CSB Technical Conference meeting, Cairns (Australia), December 2002 (Proceedings).

Palmer, T. N. and S. Tibaldi, 1988. On the prediction of forecast skill. Workshop on predictability in the medium and extended range, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, 263-309).

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tribbia, 1993: Ensemble prediction. Seminar on Validation of Models over Europe, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, vol. 1, 21-66).

Petroliagis, T., R. Buizza, A. Lanziger and T. N. Palmer, 1997: Potential use of the ECMWF ensemble prediction system in cases of extreme weather events. *Meteor. Appl.*, **4**, 69-84.

Poincaré, H., 1914. Science et Méthode. Flammarion, Paris, 71-74.

Popper, K. R., 1982. The open universe. Hutchinson, London, 6-12.

Rabier, F., E. Klinker, P. Courtier et A. Hollingsworth, 1996. Sensitivity of forecast errors to initial conditions. *Quart. J. Roy. Meteor. Soc.*, **122**, 121-150.

Richardson, D. S, 2000. Skill and economic value of the ECMWF Ensemble Prediction System, *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.

Richardson, D. S., 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473-2489.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191-201.

Szunyogh, I. & Z. Toth, 2002. The effect of increased horizontal resolution on the NCEP global ensemble mean forecasts. *Mon. Wea. Rev.*, **130**, 1125-1143.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts. Proceedings, 1-25.

Tennekes, H., 1991. Karl Popper and the accountability of numerical forecasting. New developments in predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts. Proceedings, 21-27.

Tian, H., X. Yang and X. Huangfu, 2002. Operational ensemble prediction system at China National Meteorological Center. WMO/CSB Technical Conference meeting, Cairns (Australia), December 2002 (Proceedings).

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

Toth, Z., Y. Zhu and T. Marchok, 2000: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 463-477.

Tracton M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting*, **8**, 379-398.

Whitaker, J. S. & A. F. Loughe, 1998. The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292-3302.

Wilson, L. J., W. R. Burrows and A. Lanzinger, 1999. A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127**, 956-970.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson and K. Mylne, 2001: The economic value of ensemble based weather forecasts. *Bull. Amer. Meteorol. Soc.*, **83**, 73-83.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52**, 280-299.

Historique des travaux réalisés

J'ai débuté ma carrière de météorologiste en août 1986, au service central de Météo-France, à Paris. Pendant près de 10 années j'ai exercé le dur métier de prévisionniste, à Paris puis à Toulouse mais aussi "sur le terrain" à l'occasion de fréquentes assistances, accumulant l'expérience, les prévisions réussies et les échecs. J'ai été nommé chef prévisionniste en 1995, mais je n'ai jamais occupé ce poste car j'ai quitté la France au même moment pour rejoindre la section des opérations météorologiques du CEPMMT, alors dirigée par Bernard Strauss.

Dès le début des années 1990 je m'étais intéressé de près aux premiers essais de prévision d'ensemble menés au CEPMMT (Atger, 1993 ; Atger & Lefort, 1993). A partir de 1993 j'ai travaillé activement à faire évoluer les pratiques de la prévision à moyenne échéance à Météo-France, en m'appuyant sur les résultats des contrôles subjectifs effectués en routine par les prévisionnistes (Atger, 1994b). C'est aussi à cette époque que j'ai découvert, et adopté, les positions défendues par Anders Persson (alors au CEPMMT) quant aux techniques d'interprétation pour la prévision du temps à moyenne échéance. En 1994 j'ai mis en place à Toulouse, avec quelques collègues, une méthode expérimentale de prévision qui s'appuyait à la fois sur l'utilisation du modèle à haute résolution du CEPMMT (alors en T213) et sur celle du premier ensemble opérationnel (alors en T63). C'est à la suite de la présentation des résultats de cette expérience au CEPMMT (Atger, 1995a ; Atger, 1995b) que j'ai été recruté pour "évaluer les produits du système de prévision d'ensemble, travailler sur le développement de nouveaux produits, étudier et développer une méthodologie pour l'utilisation combinée du modèle déterministe et de la prévision d'ensemble".

Pendant mon séjour au CEPMMT j'ai d'abord beaucoup travaillé sur les aspects méthodologiques (Atger, 1996). J'ai mis au point une technique de classification des membres d'un ensemble, le *tubing*, alternative à la méthode dite de *clustering* qui était alors la seule utilisée (Atger, 1997a ; Atger, 1997b). La

technique consiste en l'identification de cylindres dans un hyper-espace constitué par les membres de l'ensemble (Atger, 1999b). Le tubing privilégie à la fois la zone "centrale" de l'ensemble où la densité de membres est élevée, qui représente l'évolution météorologique la plus probable, et les "extrêmes" de l'ensemble, c'est à dire les membres qui expriment de manière typique les déviations possibles par rapport à l'option la plus probable. Le tubing est aujourd'hui une des méthodes opérationnelles de traitement des données issues de l'ensemble du CEPMMT. La mise au point de cette méthode de classification fut le résultat d'une réflexion globale sur l'interprétation des produits de la prévision numérique. Cette réflexion a fait l'objet d'un article de synthèse paru dans la revue La Météorologie (Atger, 2000b).

Parallèlement à ces travaux sur la classification, je me suis intéressé à la vérification de certains aspects de la performance de l'ensemble du CEPMMT. J'ai évalué par exemple l'aptitude de l'ensemble du CEPMMT à détecter les cyclogénèses méditerranéennes (Atger, 1997c). En septembre 1997, j'ai présenté les résultats d'une évaluation de prévisions ponctuelles de la distribution de probabilité du géopotentiel à 500 hPa à une commission du comité scientifique du CEPMMT chargée d'étudier les moyens d'évaluer la qualité de l'EPS. Les résultats furent également présentés à la conférence de Whistler sur les applications des statistiques à la climatologie (Atger, 1998). Ce travail a conduit à la parution d'un article (Atger, 1999a) dont les deux principales conclusions sont les suivantes : (i) Comparée à une distribution construite empiriquement à partir de la seule prévision de contrôle et de statistiques d'erreur du modèle, la distribution d'un ensemble n'est véritablement performante qu'à partir du quatrième jour d'intégration; (ii) Une distribution construite à partir de la moyenne des runs déterministes de quelques centres de prévision numérique est plus performante que la distribution d'un ensemble jusqu'au sixième jour d'intégration au moins. Les autres résultats portent sur la comparaison entre les ensembles opérationnels du CEPMMT et du NCEP, ainsi que sur l'impact du

nombre de membres de l'ensemble sur sa performance. Cet article constitue l'essentiel du premier chapitre de cette thèse.

J'ai terminé mon séjour au CEPMMT en mars 1998. Après une interruption de plusieurs mois, en raison de mon affectation sur un poste de prévisionniste à Toulouse, j'ai repris mes travaux de recherche en 1999 par une étude sur la prévisibilité des précipitations extrêmes. Il s'agissait initialement de répondre à une demande du CEPMMT, à la suite de plusieurs épisodes de fortes pluies survenus en 1996-1998 et ayant causé des inondations spectaculaires dans différents pays européens. Je m'étais déjà intéressé à la prévision des précipitations plusieurs années auparavant (Atger, 1994a) et j'avais acquis la conviction que la forte variabilité spatiale des précipitations impose une approche probabiliste. Le travail réalisé a été présenté à l'assemblée générale 2000 de l'EGS (European Geophysical Society) (Atger, 2000a) et a conduit à la parution d'un article en 2001 (Atger, 2001). Il s'agit d'une estimation de la valeur économique potentielle de prévisions à moyenne échéance de fortes précipitations sur la France, utilisant un modèle simple de décision. Le risque de dépassement de seuil est évalué à partir de la distribution spatiale des précipitations prévues par l'intégration unique d'un modèle ou par un ensemble. L'étude confirme la légère supériorité de l'ensemble du CEPMMT par rapport à un run unique, même à plus haute résolution. L'ensemble du NCEP est nettement moins performant à cet égard. Le second chapitre de cette thèse s'articule autour de cet article.

En 1999, j'avais présenté à l'assemblée générale de l'EGS les résultats d'un travail préliminaire sur la relation entre la dispersion d'un ensemble et la fiabilité d'une prévision déterministe (la fameuse "spread-skill relationship") (Atger, 1999d). L'objectif de cette étude, conduite pour l'essentiel pendant les dernières semaines passées au CEPMMT en 1998, était de dénoncer les méthodes utilisées par certains chercheurs pour mettre en évidence de manière déterministe le lien entre dispersion et performance. Si ce lien existe (ce qui est souhaitable) sa force ne peut être évaluée que par une approche strictement

probabiliste, du même type que celle utilisée pour évaluer la performance des prévisions probabilistes issues d'un ensemble. J'ai plusieurs fois repris des bribes de ce travail, en particulier à l'occasion d'une évaluation (non publiée) de la performance de l'indice de confiance diffusé par Météo-France, depuis 1998, pour caractériser l'incertitude de la prévision à moyenne échéance.

Durant l'été 2000, à l'occasion d'un bref séjour au CEPMMT, j'ai commencé une étude sur la variabilité spatiale et temporelle de la fiabilité de l'ensemble. Je m'étais déjà intéressé aux effets de cette variabilité sur l'estimation de la performance, mais également aux conséquences pour la calibration de prévisions probabilistes (Atger, 1999c). Ce travail s'est poursuivi en 2001 et a donné lieu à une présentation à l'assemblée générale de l'EGS (Atger, 2001b) puis à un article accepté en 2002 (Atger, 2003a). La fiabilité de prévisions probabilistes d'anomalies modérées de température à 850 hPa est évaluée sur l'Europe pendant 4 hivers consécutifs. La fiabilité locale, évaluée en chaque point considéré, est plus faible que la fiabilité évaluée sur le domaine, et très variable d'un point à un autre. La fiabilité subit également de fortes variations d'une année à l'autre. Ces variations peuvent être reliées à la variabilité spatiale et temporelle du biais systématique du modèle. Les possibilités de calibration et de correction du biais sont explorées. Cet article fait l'objet du troisième chapitre de cette thèse.

Le caractère arbitraire de la catégorisation des probabilités prévues pour la décomposition du score de Brier est un problème sur lequel je me suis penché à plusieurs reprises entre 1999 et 2002. Cette catégorisation est généralement nécessaire, en raison de la taille limitée des échantillons de vérification qui sont disponibles, mais elle conduit parfois à des estimations peu crédibles de la performance de l'ensemble. Ceci est particulièrement vrai lorsqu'on s'intéresse à la fiabilité des prévisions de fortes probabilités, dans le cas d'événements peu fréquents pour lesquels ces probabilités sont rarement prévues. J'ai proposé d'utiliser un test statistique pour réduire le caractère arbitraire de la

catégorisation. J'ai également travaillé à mettre au point une méthode de paramétrisation pour estimer la fiabilité de prévisions probabilistes issues d'un ensemble à partir d'échantillons de taille limitée. Ce travail a donné lieu à une présentation à l'assemblée générale de l'EGS en 2002 (Atger, 2002) et fait l'objet d'un article soumis pour publication en 2003 (Atger, 2003b). C'est autour de cet article qu'est rédigé le quatrième chapitre de cette thèse.

Bibliographie

(*la plupart des travaux référencés ci-dessous sont disponibles depuis la page :* http://www.multimania.com/fredatger)

Atger, F., 1993 : Current practices and perspectives in medium-range forecasting. 4th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 195-199).

Atger, F. & T. Lefort, 1993 : ECMWF Ensemble Forecast preliminary results. Expert Meeting on Ensemble Prediction System, Reading, UK, ECMWF (Report, 23-30).

Atger, F., 1994a : About quantitative precipitation prediction. Atmospheric physics and dynamics in the analysis and prognosis of precipitation fields, Rome, Italy, AGI/SIMA (Proceedings, 276-280).

Atger, F., 1994b : La fiabilité des prévisions météorologiques à moyenne échéance. *Met Mar*, **162,** 14-17.

Atger, F., 1995a : Combined use of ECMWF Ensemble Prediction System and high resolution model guidance in operational medium range forecasting. 5th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 195-199).

Atger, F., 1995b : Medium-range forecasting with EPS and T213. Expert Meeting on Ensemble Prediction System, Reading, UK, ECMWF (Report, 3-6).

Atger, F. & B. Mornet, 1995 : Operational Medium-range Weather Forecasting. 2nd European Conference on Applications of Meteorology, Toulouse, France, Météo-France and Société Météorologique de France (Proceedings, 220-224).

Atger, F., 1996 : Use of ensemble prediction in operational forecasting. Expert Meeting on Ensemble Prediction System, Reading, UK, ECMWF (Report, 37-53).

Atger, F., 1997a : Medium-range forecasting with ensemble prediction products. 6th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 216-227).

Atger, F., 1997b : The tubing, an alternative to clustering for EPS classification. Expert Meeting on Ensemble Prediction System, Reading, UK, ECMWF (Report, 31-45).

Atger, F., 1997c : Medium-range forecasting of Mediterranean cyclones. INM/WMO International Symposium on cyclones and hazardous weather in the Mediterranean, Palma de Mallorca, Spain, April 14-17 (Proceedings, 729-737).

Atger, F., 1998 : Reliability and resolution of probabilistic forecasts based on Ensemble Prediction Systems. 7th International Meeting on Statistical Climatology, Whistler, British Columbia, Canada, May 25-29.

Atger, F., 1999a : The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 9, 1941-1953.

Atger, F., 1999b: Tubing: an alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting*, **14**, 5, 741-757.

Atger, F., 1999c : Probabilistic forecasting at Météo-France. 7th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 53-56).

Atger, F., 1999d : Verification of forecasts of the forecast skill based on an ensemble prediction system. European Geophysical Society XXIV General Assembly, The Hague (The Netherlands), April 1999.

Atger, F., 2000a : Verification of intense precipitation forecasts from single models and ensemble prediction systems. European Geophysical Society XXV General Assembly, Nice (France), April 2000.

Atger, F., 2000b : La prévision du temps à moyenne échéance en France. *La Météorologie*, **8**, 30, 61-86.

Atger, F., 2001a : Performance and usefulness of ensemble based probabilistic forecasts. 8th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings,).

Atger, F., 2001b : The spatial variability of the performance of Ensemble Prediction Systems. European Geophysical Society XXVI General Assembly, Nice (France), March 2001.

Atger, F., 2001c : Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes in Geophysics*, **8**, 401-417.

Atger, F., 2002 : Estimation of the performance of ensemble based probabilistic forecasts. European Geophysical Society XXVII General Assembly, Nice (France), April 2002.

Atger, F., 2003a : Spatial and interannual variability of the reliability of ensemble based probabilistic forecasts - Consequences for calibration. *Mon. Wea. Rev.*, accepté.

Atger, F., 2003b : Estimation of the expected reliability of ensemble based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, soumis.

Chapitre 1

Performance de prévisions probabilistes basées sur des distributions construites à partir de différents modèles et systèmes de prévision d'ensemble

Autour de l'article : "The skill of ensemble prediction systems" (Atger, 1999a)

1. Introduction

Le travail décrit dans ce chapitre est né d'une volonté de relativiser les bons résultats obtenus par l'ensemble opérationnel du CEPMMT (par exemple : Buizza, 1997). Il doit beaucoup aux discussions avec Bernard Strauss, responsable de la section des "Opérations Météorologiques" du CEPMMT dans laquelle j'étais consultant entre 1995 et 1998. Notre intuition était qu'une part importante de la performance de l'EPS pouvait être obtenue par des méthodes simples et peu coûteuses, par exemple en générant des ensembles à partir d'une intégration unique du modèle. L'objectif d'un ensemble est de fournir une estimation de l'incertitude de la prévision. Une part de cette incertitude est connue *a priori*, ne serait-ce que par les statistiques d'erreur de prévision du modèle considéré. Seul l'excédent d'information doit être pris en compte pour mesurer la qualité intrinsèque d'un ensemble.

Cette approche est d'autant plus pertinente que l'objectif de la validation est de comparer des ensembles basés sur des modèles différents. En 1996-1997 un enjeu important était en effet de mettre en évidence la performance relative de l'ensemble du CEPMMT par rapport à celui du NCEP. Le débat sous-jacent était celui de la méthode de génération des perturbations initiales, les européens ayant beaucoup misé sur les (coûteux) vecteurs singuliers (Molteni et al., 1996) tandis que la méthode dite du breeding, peu onéreuse, avait la faveur des Américains (Toth et Kalnay, 1997). La querelle s'est atténuée depuis, sans complètement disparaître, du fait de l'émergence d'approches qui visent à échantillonner

l'incertitude de l'état initial en tenant compte à la fois des incertitudes relatives aux observations mais aussi des erreurs apportées par leur assimilation par un modèle imparfait (Houtekamer et al., 1996).

Un autre aspect important du travail présenté dans cet article est l'utilisation systématique de la décomposition du score de Brier (1950), proposée par Allan Murphy et plusieurs fois remaniée pendant les années 1960-70 (Murphy, 1973). Cette décomposition répondait au souci de son auteur de distinguer les deux aspects essentiels et complémentaires de la qualité d'une prévision probabiliste : d'une part la fiabilité, qui indique la correspondance entre probabilité prévue et fréquence observée de l'événement, d'autre part la résolution, qui mesure l'aptitude à prévoir des probabilités différentes suivant que l'événement prévu se réalise ou non (par exemple : Murphy, 1993). Tandis que la fiabilité indique le biais de la probabilité prévue, la résolution dépend de la capacité du système de prévision à discriminer les cas d'occurrence et de non occurrence du phénomène prévu. La pratique de la vérification montre que l'interprétation du score de Brier est souvent périlleuse si on ne tient pas compte de ces deux aspects complémentaires, en particulier lorsque ce score est utilisé pour comparer des ensembles basés sur des modèles différents.

Les premières validations des ensembles opérationnels reposaient largement sur l'évaluation de la moyenne d'ensemble, bien plus que sur la vérification de prévisions probabilistes (Palmer et al., 1993). Comme l'avait indiqué Leith (1974) longtemps avant que les moyens de calcul disponibles permettent d'envisager la production de prévisions d'ensemble à partir de modèles atmosphériques, on attend de la moyenne d'ensemble qu'elle réduise l'erreur de prévision par rapport à une prévision obtenue par intégration unique du modèle. Cette attente est justifiée par le fait que la moyenne d'ensemble a pour effet de lisser les phénomènes peu prévisibles, d'échelle réduite, tandis que les structures de grande échelle qu'on retrouve dans la plupart des intégrations sont conservées. La diminution de la variance spatiale et temporelle de la prévision induit

naturellement une diminution de l'erreur quadratique moyenne, mais aussi une augmentation de la corrélation entre prévision et observation.

La performance de la moyenne d'ensemble, comparée à une prévision de contrôle, contraste singulièrement avec la faiblesse de la relation entre dispersion de l'ensemble et qualité de la prévision (Buizza, 1997). Dans les années 1995-1999 nous étions nombreux à penser, au CEPMMT et ailleurs, que l'essentiel de ce que pouvait apporter un ensemble devait être cherché dans sa moyenne. Les produits destinés aux prévisionnistes devaient donc s'appuyer principalement sur la moyenne d'ensemble, avant d'exploiter une hypothétique "spread-skill relationship" qui apparaissait bien ténue pour être d'une grande utilité pratique (Atger, 1999b). Les contributions respectives de la moyenne et de la dispersion à la performance de l'ensemble se devaient donc d'être examinées dans le cadre de cette étude. Les résultats confirment le poids prédominant de la moyenne quant à la performance de l'ensemble.

Une des questions récurrentes dans le champ de la prévision d'ensemble est celle du nombre de membres nécessaire pour atteindre un niveau de performance requis. Après 3 années d'exploitation d'un ensemble de 33 membres, le CEPMMT décidait à la fin de 1996 d'augmenter la population de l'EPS à 51 membres. La résolution du modèle sous-jacent était également augmentée de manière substantielle (de T63 à T159) mais la question restait posée de l'impact d'un nombre de membres accru, par rapport à celui d'une meilleure résolution (Buizza et al., 1997). Au CEPMMT, comme parmi les représentants des états membres, se trouvaient des partisans d'une augmentation maximale de la résolution, au détriment de la population de l'EPS, afin de permettre une modélisation plus réaliste des phénomènes d'échelle réduite, en particulier dans les zones à forte variation de l'orographie. L'argument généralement avancé, encore aujourd'hui, pour justifier de l'augmentation du nombre de membres d'un ensemble, est celui de la prévision des phénomènes rares. Seul un ensemble très peuplé permet en effet de produire, *avant calibration*, des probabilités très faibles. Cet argument

est discuté dans le chapitre 2 de cette thèse. Dans le présent chapitre, on s'est seulement intéressé à la performance d'un EPS duquel on ne retient qu'un nombre limité de membres.

Enfin, est abordée l'évaluation d'un ensemble multi-modèles, formé à partir des prévisions issues de différents centres opérationnels. Mettre en concurrence l'EPS et cet "ensemble du pauvre" (poorman ensemble) est une autre façon de tester l'hypothèse initiale qu'une part importante de la performance de l'ensemble peut être obtenue sans intégrations multiples. Les résultats, en partie inattendus, ont conduit à ce que cette partie de l'étude prenne un relief particulier. C'est en référence à ces résultats que l'article présenté ici a le plus souvent été cité par la suite dans la littérature. Rendons à César ce qui est à César : l'idée de combiner différents modèles pour construire un ensemble est redevable à une expérience menée par Balzer & Emmrich (1997) au DWD, qui les a conduit à conclure que la dispersion entre 4 modèles opérationnels différents était un meilleur prédicteur de l'erreur de prévision que la dispersion de l'EPS. Les travaux initiés au DWD ont été poursuivis par Christine Ziehmann, alors à l'université de Postdam, avec des résultats proches de ceux obtenus dans la présente étude (Ziehmann, 2000).

2. Méthode et définitions (sections 2 et 3)

La section 2 est consacrée à la description du skill-score utilisé. Celui-ci a pour particularité d'utiliser une référence non triviale, contrairement à l'usage courant qui est d'évaluer la performance par rapport à la persistance ou à la climatologie, suivant l'échéance de prévision. La référence est ici une prévision probabiliste obtenue à partir d'une distribution normale construite autour de la prévision de contrôle de l'ensemble. L'écart type de cette distribution est pris égal à l'écart type d'erreur de prévision du modèle, tel qu'il a pu être établi à partir d'une saison antérieure. La performance de la prévision probabiliste issue de l'ensemble est ainsi évaluée par rapport à une prévision issue d'une distribution basée sur le même modèle, mais dont la dispersion est elle-même triviale puisque

indépendante de la situation et connue a priori. On mesure ainsi la performance "intrinsèque" de l'ensemble, ou plus précisément l'efficacité de la stratégie ensembliste retenue.

Les définitions des différents termes de la décomposition du score de Brier sont données en section 3. On y propose également une décomposition du skill-score en deux termes de fiabilité et de résolution. Concernant la résolution, certains auteurs ont proposé de définir un "skill" comme le rapport entre le terme de résolution du score de Brier et le terme d'incertitude (ref Lalaurette?). Cette dernière définition permet de tirer avantage du fait que la résolution est bornée par l'incertitude, qu'elle compense entièrement dans le cas d'une prévision non biaisée. En revanche elle ne donne pas naissance à une décomposition simple du skill-score.

3. Principaux résultats (sections 4 à 7)

3.1. La performance de l'ensemble du CEPMMT (section 4, a et b)

Dans la sous-section (a) est examinée la performance de la prévision de référence, obtenue à partir d'une distribution normale construite autour du contrôle de l'ensemble. La fiabilité est presque équivalente à celle d'une prévision climatologique, tandis que la résolution diminue avec l'échéance et s'approche d'un skill nul à +240 heures (Fig. 3).

Le principal résultat de la sous-section (b) est que la performance de l'ensemble n'est significative, par rapport à la prévision de référence, qu'au-delà de +96h d'échéance (Fig. 4). En réponse à la critique d'un réviseur, on discute également l'impact sur les résultats du fait que la prévision de référence, qui est une forme d'adaptation statistique, suppose une connaissance de la distribution de l'erreur de prévision du modèle, connaissance qui représente un atout indéniable par rapport à l'EPS. Enfin, on montre qu'une prévision basée sur une distribution normale construite à partir des deux seuls premiers moments de la distribution de l'EPS est pratiquement aussi performante que la prévision basée sur la

distribution originale. Ce dernier résultat semble indiquer que les moments supérieurs n'apportent pas d'information significative.

3.2. Contributions de la moyenne et de la dispersion de l'ensemble (section 4, c)

Le premier résultat présenté est qu'une distribution construite en ignorant la dispersion de l'EPS permet d'obtenir des prévisions probabilistes plus fiables et de résolution équivalente à celles obtenues à partir de la distribution originale (Fig. 4). Dans cette configuration minimaliste on abandonne toute prétention à indiquer les variations d'incertitude, puisque seules sont utilisées la moyenne de l'EPS et l'écart type (constant) de la distribution de référence.

Ce dernier résultat est interprété comme une conséquence du caractère sous-dispersif de l'EPS du CEPMMT (comme de tous les autres ensembles opérationnels à ce jour). Une méthode de calibration originale est proposée, consistant à exploiter la relation statistique entre l'écart type d'erreur du contrôle et l'écart type de l'ensemble (Fig. 5). On effectue une régression linéaire dont les termes sont obtenus en ajustant les courbes de la figure 5 par la méthode des moindres carrés. Cette correction permet de rétablir la performance due à la dispersion de l'ensemble, qui reste cependant faible, en valeur relative, par rapport à celle due à la moyenne (Fig. 4).

3.3. Impact du nombre de membres de l'ensemble (section 5)

Les résultats obtenus dans cette section ne permettent pas de tirer de conclusions sur le nombre de vecteurs singuliers qu'il est souhaitable de calculer, mais seulement sur le nombre d'intégrations nécessaires à partir des états initiaux obtenus en combinant 25 vecteurs singuliers. Cette distinction est importante car c'est le calcul des vecteurs singuliers qui est coûteux, et non l'intégration des 51 membres. Cette précision vaut également pour l'étude sur l'impact du nombre de membres qui est présentée dans le chapitre 2 (section 3.2).

Le principal résultat est que la fiabilité sature rapidement avec le nombre de membres, suivant en cela le niveau de la dispersion. En revanche la résolution

augmente encore de manière significative lorsqu'on passe de 33 à 51 membres. L'augmentation de la dispersion consécutive à l'accroissement du nombre de membres semble indiquer que l'ajout de membres, pourtant obtenus à partir de combinaisons des mêmes vecteurs singuliers, ne conduit pas seulement à un meilleur échantillonnage de la distribution prévue.

3.4. Comparaison entre les ensembles du CEPMMT et du NCEP (section 6)

Les résultats présentés dans cette section sont généralement à l'avantage de l'EPS du CEPMMT. En raison de critiques insistantes de la part des réviseurs, les points de comparaison sont largement discutés, parfois même en faisant référence à des résultats obtenus par des auteurs du NCEP (Zhu et al., 1996). La meilleure fiabilité de l'EPS du CEPMMT est, au moins partiellement, la conséquence du fort manque de dispersion de l'EPS du NCEP à moyenne échéance. Ce défaut n'est pas seulement dû au nombre réduit de membres de l'EPS du NCEP. En revanche, les deux systèmes ont une résolution très similaire.

3.5. Ensemble multi-modèles (section 7)

Le principal résultat, largement commenté depuis sa publication et confirmé par plusieurs auteurs (Talagrand et al., 1997 ; Ziehmann, 2000 ; Ebert, 2001 ; Ebert, 2002), est qu'un ensemble multi-modèles, ne nécessitant pas d'intégration multiple, est plus performant que les EPS opérationnels jusqu'à l'échéance +144h environ, c'est à dire pour les échéances les plus utiles, en pratique, pour la prévision opérationnelle. Cette performance est entièrement due à une meilleure résolution, la fiabilité étant équivalente ou inférieure (suivant la manière de construire la prévision à partir de la distribution brute) à celle des EPS opérationnels.

Comme on pouvait s'y attendre au vu des résultats obtenus en réduisant la population de l'EPS du CEPMMT (section 5) seule la moyenne d'ensemble contribue à la performance, l'ensemble multi-modèles étant largement sous-dispersif en raison du nombre limité de ses membres. Cette section se termine

par un appel à expérimenter une stratégie multi-modèles et/ou multi-analyses pour la conception des ensembles. Cette approche est aujourd'hui celle de plusieurs projets, en particulier le MAED (Multi Analysis Ensemble Data) du CEPMMT et le PEPS (Poorman's Ensemble Prediction System) du Met'Office, coordonné par Ken Mylne (communication personnelle).

4. Remarques conclusives (section 8)

Le premier point discuté dans cette section est celui de l'impact du seuil choisi (50 m) sur les résultats. On montre d'une part que la résolution ne dépend pas du seuil, d'autre part que la contribution de la dispersion à la performance, par le biais de la fiabilité, devient positive au-delà d'un certain seuil. Par ailleurs, on observe que la performance relative d'un ensemble multi-modèles, par rapport à un EPS opérationnel, est inchangée lorsque le seuil passe de 50 m à 150 m. On montre ensuite que l'utilisation d'un critère de performance autre que le score de Brier ou les termes de sa décomposition, tel que l'aire sous la courbe ROC (diagnostic décrit dans le chapitre 2 de cette thèse), ne remet pas en cause les résultats obtenus.

Bibliographie

Atger, F., 1999a. The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 9, 1941-1953.

Atger, F., 1999b. Tubing: an alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting*, **14**, 5, 741-757.

Balzer, K. and P. Emmrich, 1997: Some remarks on the assessment of EPS. Expert Meeting on Ensemble Prediction System, Reading, UK, ECMWF (Report, 49-53).

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Buizza, R., 1997: Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 99-119.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, N. Wedi, 1997: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935-1960.

Ebert, E.E., 2001: Ability of a poor Man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461-2480.

Ebert, E.E., 2002: Corrigendum. *Mon. Wea. Rev.*, **130**, 1661-1663.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie and H.L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73-119.

Murphy, A., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tribbia, 1993: Ensemble prediction. Seminar on Validation of Models over Europe, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, vol. 1, 21-66).

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings of the Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts, 1-25.

Toth Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conference on Weather Analysis and Forecasting, Norfolk, Virginia, Amer. Meteor. Soc., J79-J82.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52**, 280-299.

The Skill of Ensemble Prediction Systems

FRÉDÉRIC ATGER

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 6 April 1998, in final form 29 September 1998)

ABSTRACT

The performance of ensemble prediction systems (EPSs) is investigated by examining the probability distribution of 500-hPa geopotential height over Europe. The probability score (or half Brier score) is used to evaluate the quality of probabilistic forecasts of a single binary event. The skill of an EPS is assessed by comparing its performance, in terms of the probability score, to the performance of a reference probabilistic forecast. The reference forecast is based on the control forecast of the system under consideration, using model error statistics to estimate a probability distribution. A decomposition of the skill score is applied in order to distinguish between the two main aspects of the forecast performance: reliability and resolution. The contribution of the ensemble mean and the ensemble spread to the performance of an EPS is evaluated by comparing the skill score to the skill score of a probabilistic forecast based on the EPS mean, using model error statistics to estimate a probability distribution.

The performance of the European Centre for Medium-Range Weather Forecasts (ECMWF) EPS is reviewed. The system is skillful (with respect to the reference forecast) from +96 h onward. There is some skill from +48 h in terms of reliability. The performance comes mainly from the contribution of the ensemble mean. The contribution of the ensemble spread is slightly negative, but becomes positive after a calibration of the EPS standard deviation. The calibration improves predominantly the reliability contribution to the skill score. The calibrated EPS is skillful from +72 h onward.

The impact of ensemble size on the performance of an EPS is also investigated. The skill score of the ECMWF EPS decreases steadily with reducing numbers of ensemble members and the resolution is particularly affected. The impact is mainly due to the ensemble spread contributing negatively to the skill. The ensemble mean contribution to the skill decreases marginally when reducing the ensemble size up to 11 members.

The performance of the U.S. National Centers for Environmental Prediction (NCEP) EPS is also reviewed. The NCEP EPS has a lower skill score (vs a reference forecast based on its control forecast) than the ECMWF EPS especially in terms of reliability. This is mainly due to the smaller spread of the NCEP EPS contributing negatively to the skill. On the other hand, the NCEP and ECMWF ensemble means contribute similarly to the skill. As a consequence, the performance of the two systems in terms of resolution is comparable.

The performance of a poor man's EPS, consisting of the forecasts of different NWP centers, is discussed. The poor man's EPS is more skillful than either the ECMWF EPS or the NCEP EPS up to +144 h, despite a negative contribution of the spread to the skill score. The higher skill of the poor man's EPS is mainly due to a better resolution.

1. Introduction

Until the advent of ensemble prediction systems (EPSs) the evaluation of the quality of numerical weather forecasts was essentially based on a space–time comparison between forecast and verifying values, with only one forecast value and one verification value occurring at the same time and the same place. Since December 1992, both the U.S. National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF) have produced operational forecasts based on ensemble prediction (Tracton and Kalnay 1993; Palmer et al. 1993). An EPS is a prediction system designed to provide an ensemble of N forecasts of the meteorological state, considered as N independent realizations of a predicted probability distribution. The quality evaluation of an EPS should thus be based on the verification of a probability distribution. This implies that the forecast error cannot be estimated from a simple comparison between a forecast value and a verifying value. While the forecast is a distribution of values, the basic verification is still a single value. Two different approaches can be followed to solve this dilemma:

- The quality of a single probability distribution forecast (one time, one location) is estimated from the conditional probability that the actual verification occurs, given the probability distribution (Wilson 1995). In this Bayesian approach, the performance depends on two independent aspects: (i) how close the distribution

Corresponding author address: Dr. Frédéric Atger, Météo-France (SCEM/PREVI), 42, av. G. Coriolis, 31057 Toulouse Cedex, France. E-mail: frederic.atger@meteo.fr

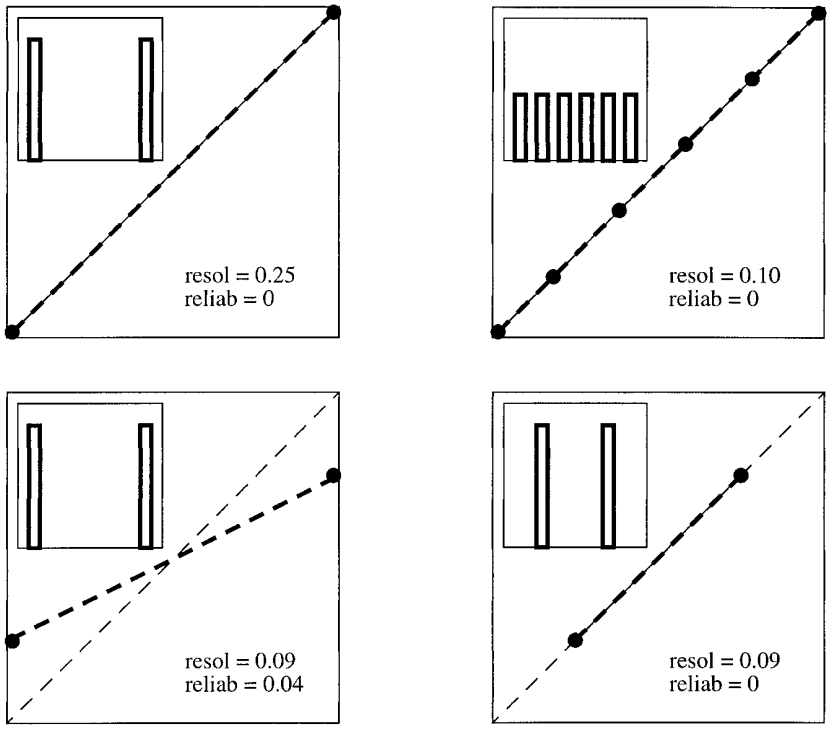


FIG. 1. Typical *reliability* diagrams (showing predicted probability vs observed frequency) and *sharpness* histograms (showing the distribution of predicted probabilities). (a) Perfect resolution and reliability, perfect sharpness. (b) Perfect reliability but poor sharpness, lower resolution than (a). (c) Perfect sharpness but poor reliability, lower resolution than (a). (d) As in (c) but after calibration, perfect reliability, same resolution.

mode is to the verifying value and (ii) how sharp is the distribution.

- The quality of a single forecast is not estimated per se, but a set of probability distribution forecasts is compared with the distribution of the corresponding verification values. The performance depends again on two aspects: (i) the correspondence between the predicted probability and the actual frequency of occurrence and (ii) the prominence of lower and higher probabilities, which are more meaningful than probabilities close to the climatological frequency.

The more common second approach has been adopted in this study of the quality of ensemble prediction systems. The performance has been assessed by comparing the distribution of ensemble forecasts to a reference distribution based on the control forecast of the considered EPS. A skill score resulting from this approach is defined in section 2. A decomposition of the skill score, according to the two independent aspects of the performance mentioned above, is proposed in section 3. The performance of the ECMWF EPS is reviewed in section 4. The question of the impact of reducing the ensemble size is addressed in section 5. The ECMWF EPS and the NCEP EPS are compared in section 6. A poor man's scheme is used in section 7 to investigate the cost efficiency of an EPS. The main results of the study are summarized and discussed in section 8.

2. Skill score

A skill score measures the ability of a forecast to obtain a better score than a reference forecast. Skill scores are usually computed using either the climatology or the persistence of initial conditions as a reference forecast. As far as the performance of an EPS is concerned, the skill versus the climatology indicates the overall performance of the system (e.g., Buizza 1997) but does not allow one to distinguish the relative performance of the *model* on which it is based from the performance of the EPS itself, that is, the generation of the perturbed forecasts. In this study the skill of an EPS is assessed using the single control forecast of the EPS as a reference, in order to judge the benefit of the extra ensemble members.

The score used in this study is the (half) Brier score (Brier 1950) or probability score (PS), that is, the average square difference between the forecast probability p_i and the observation o_i with $o_i = 1$ if observed, $o_i = 0$ if not observed. The number of individual forecasts is N:

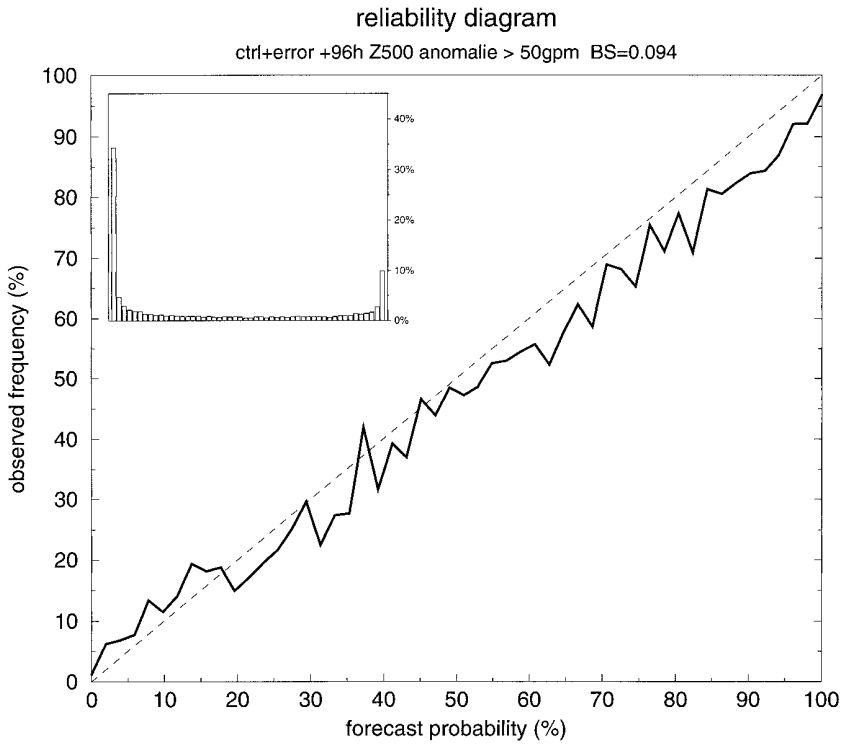


FIG. 2. Reliability diagram and sharpness histogram of the reference forecast based on the ECMWF EPS control forecast (see text for details). Winter 1996/97, Europe, 500-hPa geopotential height anomaly exceeding 50 m, +96 h.

$$\mathrm{PS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2.$$

The probability skill score (PSS) is conventionally defined as the relative probability score compared with the probability score of the reference forecast. It can be expressed in percent of the potential improvement over the reference score:

$$\mathrm{PSS} = \frac{\mathrm{PS}_{\mathrm{ref}} - \mathrm{PS}}{\mathrm{PS}_{\mathrm{ref}}}.$$

The performance of different EPSs has been assessed regarding the forecast of a simple binary event, the 500-hPa geopotential height anomaly exceeding 50 m (above +50 m or below −50 m), at 140 grid points over Europe (75°N, 20°W; 30°N, 45°E) every 5° of latitude and longitude, weighted according to the latitude. The period of verification is from 10 December 1996 to 28 February 1997 (starting when the improved, higher-resolution ECMWF EPS was implemented). The total number of individual events considered in the study (140 points × 81 days = 11 340), although not independent, is believed to be large enough to get significant results. The sensitivity of the results to the choice of the 50-m threshold is discussed in section 8.

The reference forecast is a Gaussian distribution of the 500-hPa geopotential height, with a mean equal to the EPS control forecast, and a standard deviation equal to the standard deviation of the control forecast error estimated over an independent winter season (10 December 1995–28 February 1996).

In order to get a comparison as fair as possible between the reference forecast and the EPS forecast, the same number of probability categories, depending on the number of ensemble members of the considered EPS, has been used for both systems, that is, $N + 1$ categories from $0/N$ to N/N for evaluation of an N-member EPS.

3. Skill score decomposition

The probability score decomposition proposed by Murphy (1973) has been used in order to distinguish the two main aspects of the forecast performance: statistical consistency, or *reliability*, and variability, or *resolution*. Murphy's decomposition consists of three terms:

$$\mathrm{PS} = \sum_{k=1}^{T}\frac{n_k}{N}(p_k - \overline{o}_k)^2 + \sum_{k=1}^{T}\frac{n_k}{N}(\overline{o}_k - \overline{o})^2 + \overline{o}(1 - \overline{o}),$$

when a sample of N forecasts has been divided in T categories, each comprising n_k forecasts of a probability

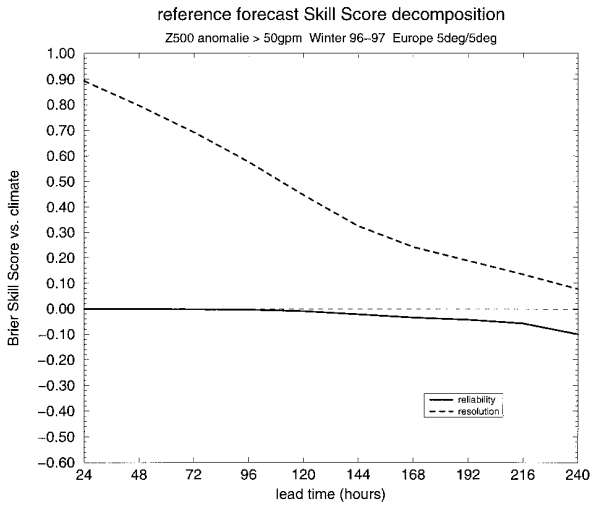


FIG. 3. Decomposition of the skill score of the reference forecast vs a forecast based on the long-term climatology. Winter 1996/97, Europe, 500-hPa geopotential height anomaly exceeding 50 m, +96 h.

p_k, o_k being the observed frequency when the forecast was lying in that category and o the observed frequency in the whole sample.

- The first term is the reliability, that is, the average square difference between the forecast probability and the observed frequency in the different categories. This term indicates the ability of the system to forecast accurate probabilities, so that for example an observed frequency of 30% can be expected when a 30% probability is forecast. The reliability is negatively oriented, as is the probability score: the lower the reliability the better. A perfect reliability (0) is indicated by a curve lying along the diagonal of a reliability diagram, that is, a plot of forecast probability versus observed frequency (Fig 1).
- The second term is the resolution, that is, the average square difference between the observed frequency in each category and the mean frequency observed in the whole sample. This term indicates the ability of the forecast to separate the different categories, whatever the forecast probabilities. For a given reliability, the resolution thus indicates the sharpness of the forecast (Fig. 1): the maximum resolution corresponds to a deterministic forecast (only 0% and 100% are forecast), the minimum resolution corresponds to a climatological forecast (only one probability is forecast). The resolution is positively oriented: the higher the resolution, the better.
- The third term is the uncertainty, that is, the variance of the observations, indicating the intrinsic difficulty in forecasting the event during the period. It is also the probability score of the sample climatology forecast. The uncertainty is obviously independent of the forecast system: being the same for the reference forecast and the forecast under evaluation, it plays no role in the skill score.

The skill score defined above can thus be decomposed into two terms, positively oriented, indicating (i) the skill due to the reliability and (ii) the skill due to the resolution:

$$\mathrm{PSS} = \left(\frac{\mathrm{RELIABILITY}_{\mathrm{ref}} - \mathrm{RELIABILITY}}{\mathrm{PS}_{\mathrm{ref}}}\right) + \left(\frac{\mathrm{RESOLUTION} - \mathrm{RESOLUTION}_{\mathrm{ref}}}{\mathrm{PS}_{\mathrm{ref}}}\right).$$

Reliability and resolution are independent. For example, if the observed frequency is 90% in the 10% probability category, and 10% in the 0% probability category, the resolution is high but the reliability is poor. For operational purposes, the resolution term is the most relevant, since the reliability, as any bias, can generally be improved by a calibration. This can be done for instance by replacing forecast probabilities by the actual frequencies observed during a previous season in the same categories (Zhu et al. 1996). The reliability improvement is obtained at the expense of *sharpness,* as illustrated by the histograms in Fig. 1 showing the distribution of predicted probabilities. The resolution is not modified by the calibration if the number of categories remains the same.

4. The skill of the ECMWF EPS

The ECMWF ensemble prediction system comprises (in its current high-resolution version implemented in December 1996) 50 perturbed and an unperturbed, control integration of a T_L159L31 version of the ECMWF model (Simmons et al. 1989; Courtier et al. 1991). The ECMWF EPS methodology is described in Molteni et al. (1996) while the advantages of the more recent, higher-resolution system are discussed in Buizza et al. (1998).

a. The reference forecast

The reference forecast is based on the control forecast as defined in section 2. The T_L159 version of the ECMWF model on which the EPS was based in winter 1996/97 was not operational in winter 1995/96. For this reason, the standard deviation of the control forecast error was estimated from the T213 ECMWF model instead, assuming that the two models are close enough to lead to the same error distribution characteristics over a season. The reference forecast is particularly good in the early medium range. The reliability diagram on +96 h (Fig. 2) exhibits a good reliability and a sharp distribution, leading to a low (i.e., good) probability score. This is not the case as the forecast range increases, when the distribution becomes flatter and the forecast less reliable.

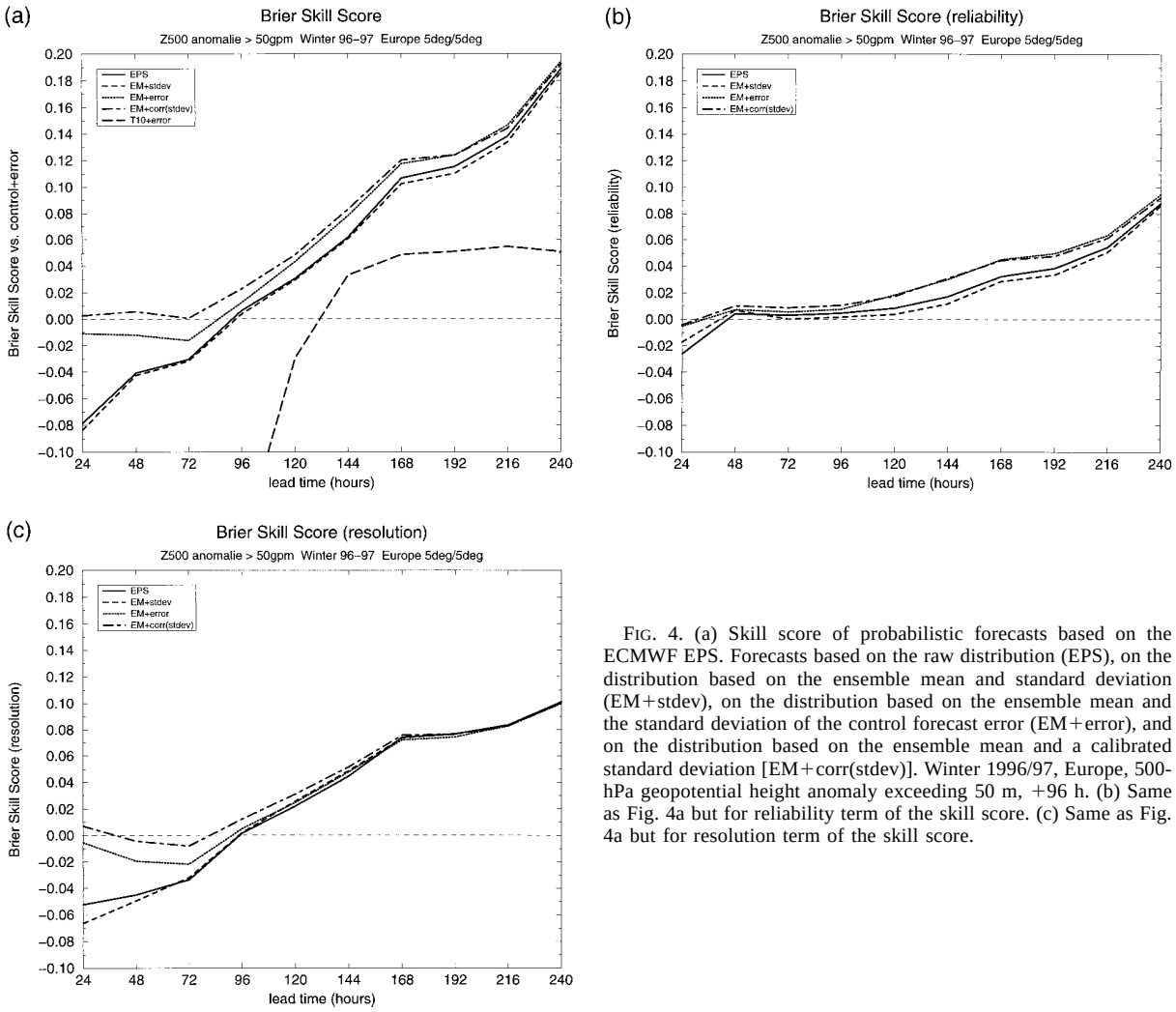


FIG. 4. (a) Skill score of probabilistic forecasts based on the ECMWF EPS. Forecasts based on the raw distribution (EPS), on the distribution based on the ensemble mean and standard deviation (EM+stdev), on the distribution based on the ensemble mean and the standard deviation of the control forecast error (EM+error), and on the distribution based on the ensemble mean and a calibrated standard deviation [EM+corr(stdev)]. Winter 1996/97, Europe, 500-hPa geopotential height anomaly exceeding 50 m, +96 h. (b) Same as Fig. 4a but for reliability term of the skill score. (c) Same as Fig. 4a but for resolution term of the skill score.

Comparing the reference forecast to a climatological forecast based on the ECMWF 15-yr reanalysis (Gibson et al. 1997), the skill score of the reference forecast has been computed then decomposed as indicated in section 3. The reference forecast has some skill versus the climatological forecast up to +240 h. The resolution is the only source of skill, while the reliability component tends to become slightly negative in the late medium range (Fig. 3). The climatological forecast has optimal reliability, since its distribution is expected to match the sample observations distribution, unless in an anomalous period. A similar behavior is observed for the reference forecast, since it is based on model error distribution. The loss of reliability in the late medium range may be due to the fact that the model error is not independent of forecast values, also to slight differences between the T213 ECMWF model used for the error statistics and the actual EPS T_L159 model.

b. The skill score of the ECMWF EPS forecast

The skill score of the ECMWF EPS forecast, with respect to the reference forecast described above, is shown in Fig. 4a (solid line, labeled EPS). The EPS forecast has a negative skill in the short range, as expected since the singular vectors have a 48-h optimization time interval (the system is designed for medium-range prediction) and starts to be skillful around +96 h. A 10% skill score is reached at +168 h, 19% at +240 h is the maximum. The decomposition of the skill score shows that the reliability component (Fig. 4b) starts to be slightly positive earlier (+48 h) than the resolution component (Fig. 4c). On the other hand the resolution component grows faster from +96 h onward.

The reference forecast used in this study is based on a certain knowledge of the model error distribution, from past verification statistics. The EPS forecast, unless

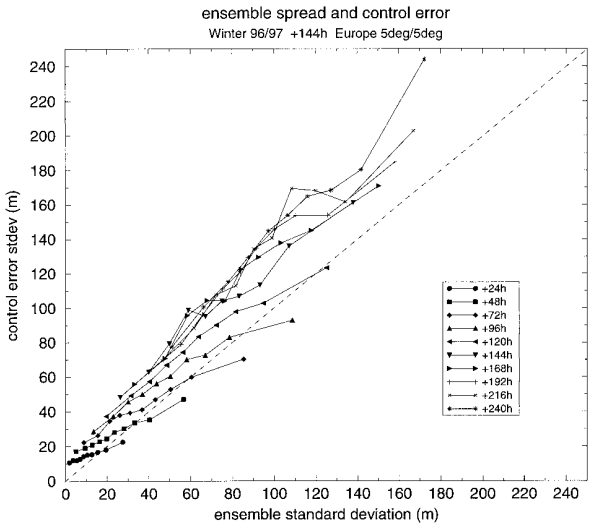


FIG. 5. Relation between the ECMWF EPS standard deviation, stratified in 10 equally populated categories, and the standard deviation of the ECMWF EPS control error. Winter 1996/97, Europe, 500-hPa geopotential height.

it is calibrated (as later on in this paper), does not assume such a knowledge, so that its comparison to the reference forecast might be considered as disadvantageous. A more neutral, less skillful reference forecast is obtained by replacing the standard deviation of the model error by the climate monthly standard deviation. This forecast is worse than the actual reference forecast *only in the short range* (not shown). This is due to its lack of reliability, the climate variability being much larger than the expected forecast uncertainty. On the other hand its resolution is virtually the same as the actual reference forecast at all lead times (not shown).

The skill score of a Gaussian distribution based on the ensemble mean and standard deviation is shown in Fig. 4a (dashed line, labeled EM+stdev). In this case, only the first two moments are used and the distribution is assumed to be monomodal and symmetric. Yet the skill score is almost exactly the same as the original distribution, especially in terms of resolution (Fig. 4c). This seems to indicate that, as far as the probability score is concerned, the main information that can be extracted from an EPS distribution is already contained in the mean and standard deviation of this distribution.

c. *Ensemble mean and ensemble spread contributions to the skill score*

Figure 4a also shows the skill score of a Gaussian distribution based only on the ensemble mean, using an estimate of the standard deviation of the control forecast error (as for the reference forecast) instead of the EPS standard deviation (dotted line, labeled EM+error). Here the ensemble standard deviation is ignored, the distribution still being assumed to be monomodal and symmetric. This forecast is slightly more skillful than the EPS forecast, suggesting that all the information that can be extracted from the EPS distribution is already contained in the mean of the distribution. The spread appears to have no impact in terms of resolution (Fig. 4c) and a negative impact in terms of reliability (Fig. 4b).

The ECMWF EPS standard deviation is well known to be too small on average compared with the control error, although they tend to be correlated (Buizza 1997). A way to highlight this ''hidden'' information is to calibrate the EPS standard deviation by taking into account this correlation, as it is known from the past. Figure 5 shows that there is indeed a linear correlation between the mean value of different categories of EPS standard deviation and the standard deviation of the control error in these categories during winter 1996/97. Assuming that this result could have been known from previous months or years of data (it is not the case since the T_L159 EPS has run only since December 1996), a forecast distribution based on the ensemble mean and the *corrected ensemble standard deviation* has been constructed. The main, positive impact of the correction is on the reliability part of the skill score, the resolution being improved only in the short range. Overall the calibrated forecast has some skill from +72 h onward and is even slightly more skillful than the distribution based on the ensemble mean and the standard deviation of the control error [Fig. 4a, dotted–dashed line, labeled EM+corr(std)].

Nevertheless the ensemble mean appears to be the main source of skill. In a recent paper, Hamill and Colucci (1998) found a similar result when assessing the performance of short-range ensemble prediction. With regard to the synoptic fields, the main benefit of the ensemble mean compared with the control forecast is to smooth out small-scale, unpredictable features, while retaining the more slowly varying large-scale pattern (Leith 1974). A similar effect is achieved by filtering high-resolution deterministic forecasts, for instance by retaining only the leading components of a spectral decomposition. The skill score of a filtered forecast, obtained by an arbitrary T10 truncation of the control forecast, positive from +144 h, is still far smaller than the skill score of a forecast based on the ensemble mean and the standard deviation of the control error (Fig. 4a, long dashed line, labeled T10+error).

The comparatively poor contribution of the ensemble spread to the skill, even after calibration, might be related to an insufficient number of ensemble members causing poorer sampling of the higher moments of the distribution. However, it should be stressed that the good performance of the ensemble mean is obviously due to the ensemble spread being reliable enough to ''shift'' the mean of the distribution from the control forecast. The decomposition proposed in this section allows an assessment of the relative importance of reliable information already contained in the ensemble mean, versus

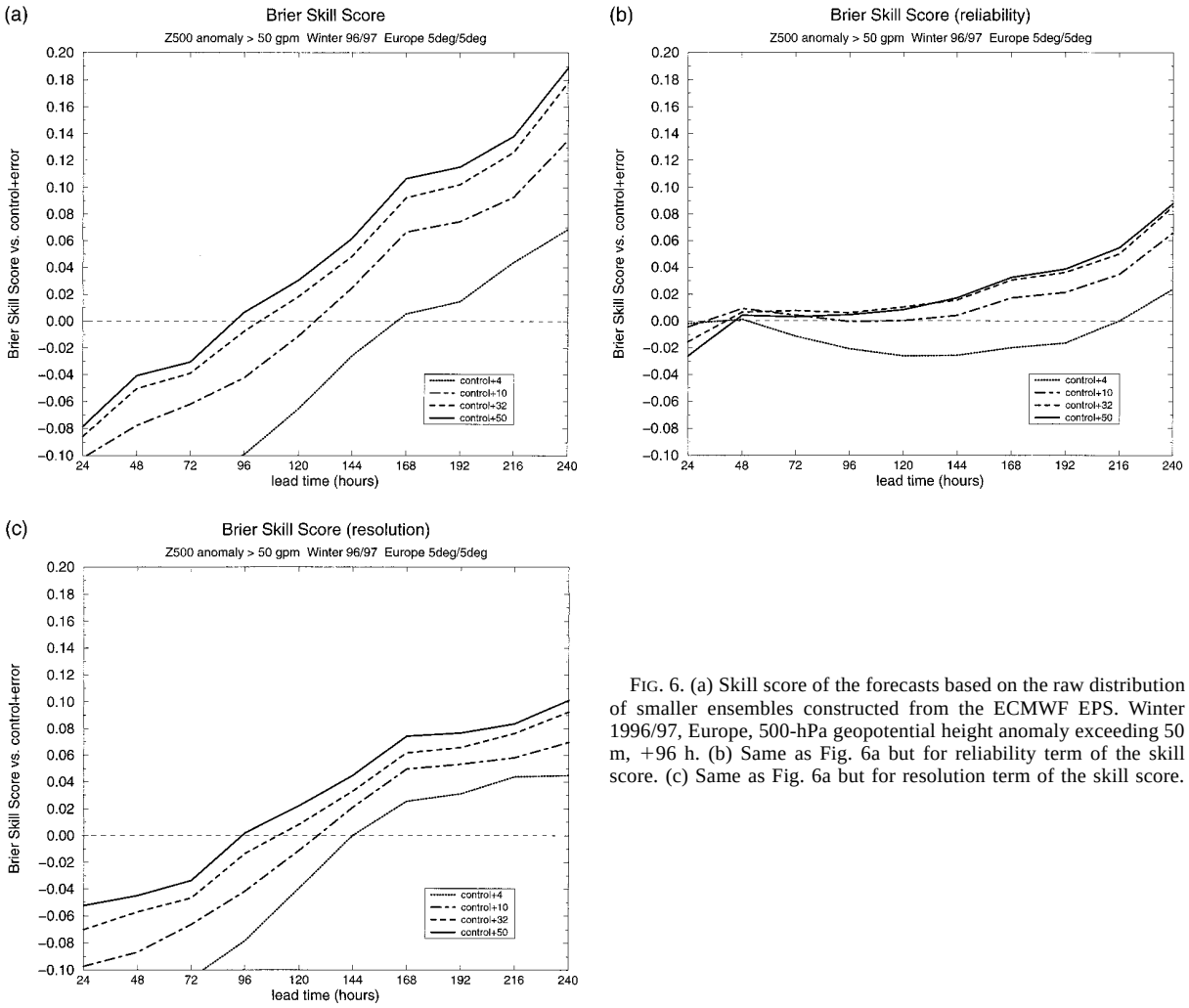


FIG. 6. (a) Skill score of the forecasts based on the raw distribution of smaller ensembles constructed from the ECMWF EPS. Winter 1996/97, Europe, 500-hPa geopotential height anomaly exceeding 50 m, +96 h. (b) Same as Fig. 6a but for reliability term of the skill score. (c) Same as Fig. 6a but for resolution term of the skill score.

the residual information that can be found in the spread itself. Therefore, the results presented here are not in contradiction with several studies showing the value of ensemble spread in forecasting the forecast skill (e.g., Buizza 1997; Toth et al. 1998)

5. Smaller ensembles

This section addresses the issue of the impact of ensemble size on the skill score. *Smaller* ensembles, constructed using the ECMWF EPS control and the 4, 10, and 32 first perturbed members as in Buizza and Palmer (1998), were compared with the whole ensemble (control + 50 members) in terms of skill score. Note that the first perturbed members initial conditions are still obtained from *all* 25 singular vectors (SVs) (since the perturbations are combinations of SVs). This comparison thus favors smaller ensembles and only addresses the question of the number of integrations that are needed. As in the previous section the skill score has been computed for three different distributions: (i) the raw EPS distribution, (ii) a Gaussian distribution based on the ensemble mean and standard deviation, and (iii) a Gaussian distribution based on the ensemble mean and the standard deviation of the control forecast error.

As expected, the skill score of the raw distribution decreases when the EPS population is reduced (Fig. 6a). The performance is almost the same from 33 to 51 members, then decreases rapidly. These results are similar to those obtained recently by Talagrand et al. (1998) for 850-hPa temperature probabilistic forecasts. Resolution and reliability behave differently in that respect. Reliability is almost not affected from 11 to 51 members, especially in the shorter ranges (Fig. 6b). Resolution decreases more steadily from 51 to 5 members, especially in the shorter ranges (Fig. 6c).

The skill score of the distribution based on the ensemble mean and standard deviation tends to be larger than for the raw distribution when reducing ensemble membership up to 11 and 5 members (Fig. 7). The prob-

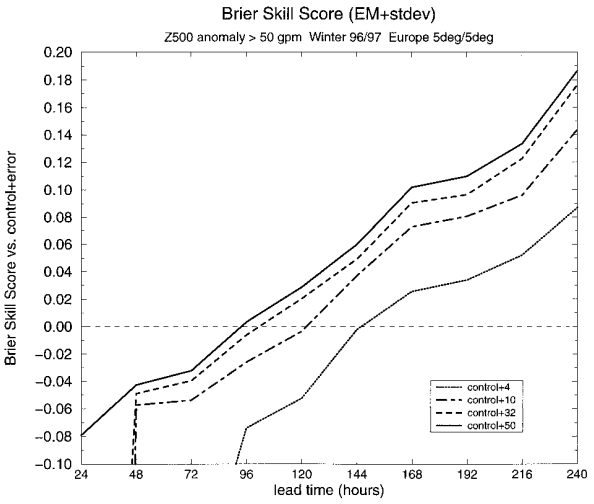


FIG. 7. Same as Fig. 6a but for distribution based on the ensemble mean and standard deviation, instead of the raw ensemble distribution.

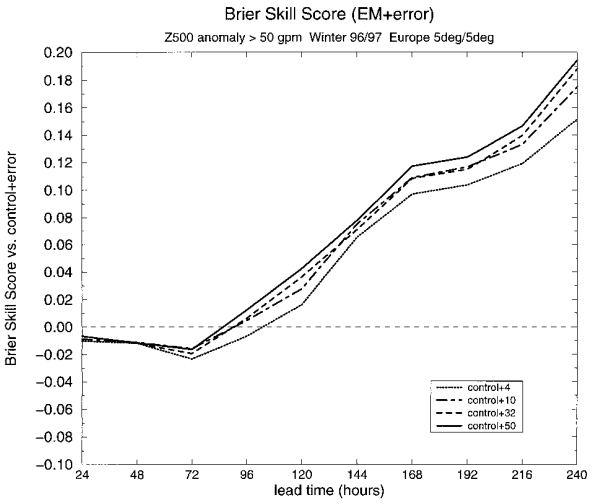


FIG. 8. Same as Fig. 6a but for distribution based on the ensemble mean and the standard deviation of the control forecast error, instead of the raw distribution.

ability score depends on the number of forecast categories: any increase of the number of realizations extracted from the same statistical population would have the effect of smoothing noise due to the finiteness of the sample. In the case of the distribution based on ensemble mean and standard deviation, the number of categories is arbitrarily fixed to 52, as for the reference forecast, whereas the number of categories is obviously limited to the number of members when considering the raw distribution.

The decrease of the skill score when reducing the ensemble size is limited when considering the distribution based on the ensemble mean and the standard deviation of the control forecast error. The skill score is significantly reduced only for a five-member ensemble (Fig. 8). The smaller ensemble mean is obviously similar to the whole ensemble mean, so that the positive impact on skill when increasing the EPS population is only due to the ensemble spread being significantly better.

Increasing ensemble size has a definite impact on the average amplitude of the spread. As shown in Table 1, the spread increases by approximately 10%–16% from 5 to 11 members, and by 6%–9% from 11 to 33 members extracted from the same subset. These figures are larger than expected from theory. The standard deviation, STD, of an ensemble of N elements randomly extracted from a distribution of standard deviation std is given by

$$\mathrm{STD} = \mathrm{std}\left(1 - \frac{1}{N}\right)^{1/2},$$

which is an increase of 7% from 5 to 11 members, and 3% from 11 to 33 members. The observed values indicate that extra members do not belong exactly to the same statistical population as the first members. Increasing the size of an EPS leads not only to a better sampling, but also to a significant improvement of the distribution.

The larger spread explains the better reliability of larger ensembles (Fig. 6b). But the resolution component of the skill score is also improved when increasing ensemble size (Fig. 6c). The resolution does not depend on the average amplitude of the spread, rather on its daily variations (see section 3). This variability seems to be improved by increasing ensemble size.

6. Comparison of ECMWF EPS with NCEP EPS

In this section the skill score has been used in order to compare two different operational ensemble prediction systems: the ECMWF EPS (50 + 1 members) and the NCEP 0000 UTC EPS (10 + 1 members). The reference forecast has been separately constructed from each EPS's own control forecast, in order to limit the performance dependence to the model on which the system is based. As in previous sections the skill score was computed for three different distributions: (i) the raw EPS distribution, (ii) a Gaussian distribution based on the ensemble mean and standard deviation, and (iii) a Gaussian distribution based on the ensemble mean and

TABLE 1. Ensemble standard deviation of 500-hPa geopotential height over Europe, winter 1996/97, ECMWF EPS, and NCEP EPS, for various memberships including the control forecast and a number of perturbed forecasts (m).

ECMWF (NCEP)	(4 + 1) members	(10 + 1) members	(32 + 1) members	(50 + 1) members
+ 96 h	37 m (30 m)	43 m (33 m)	47 m	47 m
+144 h	55 m (43 m)	63 m (49 m)	68 m	69 m
+240 h	83 m (67 m)	92 m (75 m)	98 m	99 m

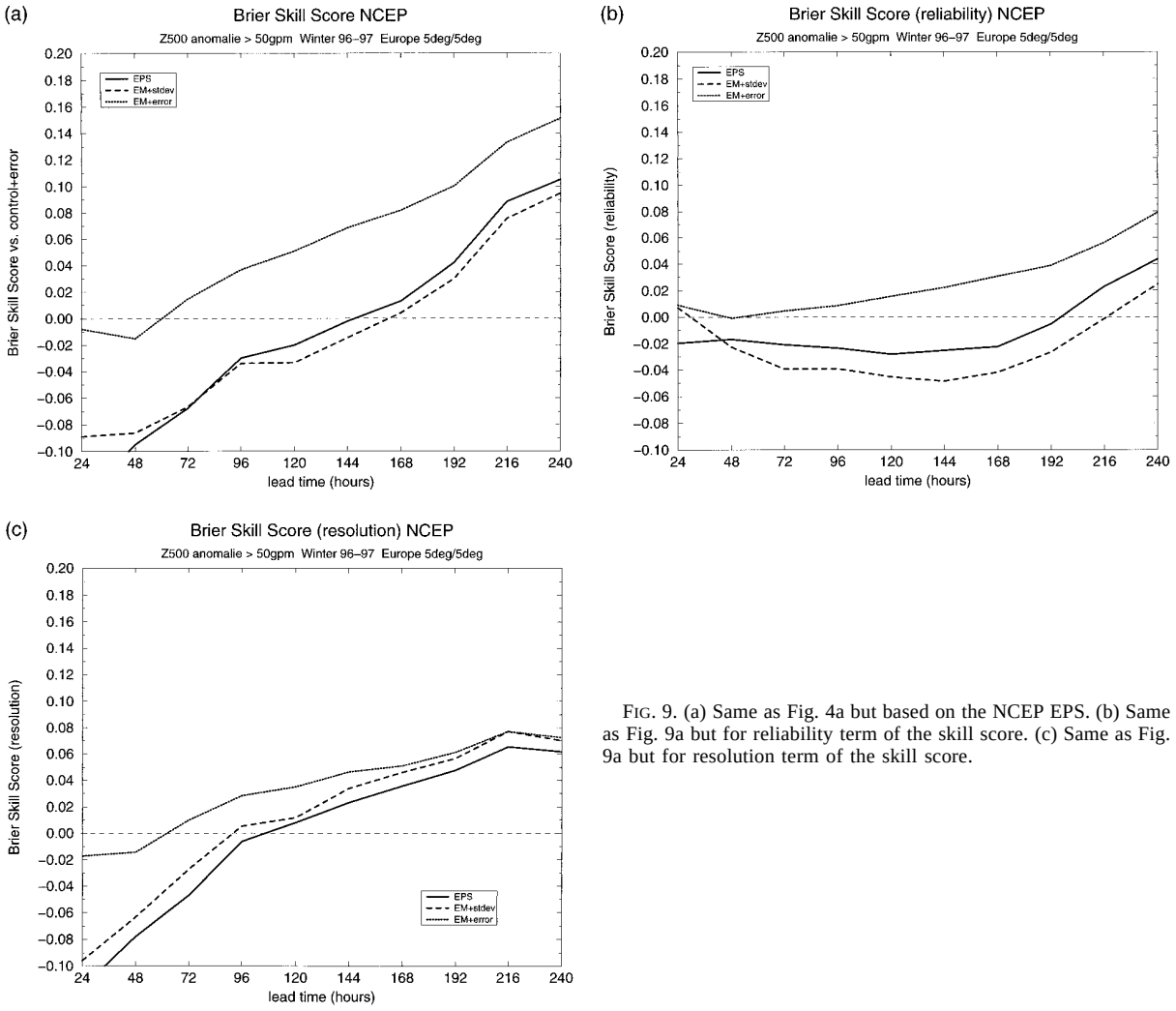


FIG. 9. (a) Same as Fig. 4a but based on the NCEP EPS. (b) Same as Fig. 9a but for reliability term of the skill score. (c) Same as Fig. 9a but for resolution term of the skill score.

the standard deviation of the control forecast error. When it was possible (ii and iii) the number of categories used for the computations was the same (52) for both systems in order to get a comparison as fair as possible. The same analysis (ECMWF analysis) was used to verify both systems. Using NCEP analysis to verify NCEP EPS might have had a slight, positive impact on NCEP skill. This was not possible at the time of the study, for technical reasons.

The NCEP ensemble prediction system based on 0000 UTC consists of 10 perturbed and 1 unperturbed control integrations of a T62 version of the NCEP model. The main difference from the ECMWF EPS, besides the number of members, lies in the way the initial perturbations are generated. While the singular vectors technique is used at ECMWF (Molteni et al. 1996), the NCEP perturbations are obtained through the method of breeding of growing modes (Toth and Kalnay 1997).

The NCEP EPS proves less skillful overall than the ECMWF EPS (cf. Fig. 9a and Fig. 4a). The raw distribution starts to be skillful at +144 h only (+96 h for the ECMWF EPS) and reaches a maximum of 10% at +240 h (19% for the ECMWF EPS). The distribution based on the ensemble mean and standard deviation is even worse.

The performance of the distribution based on the ensemble mean and the standard deviation of the control forecast error is rather similar for the two systems (cf. Fig. 9a and Fig. 4a). NCEP EPS is slightly better in the early medium range, while ECMWF EPS is slightly better in the late medium range. A similar behavior was observed by Zhu et al. (1996) in terms of different verification statistics for an independent winter season. The differences between the two systems at earlier lead times might be related to differences in initial perturbations. One would expect in this case a better performance of NCEP EPS as early as +24 h. This is not observed. At later lead times the differences in model performances

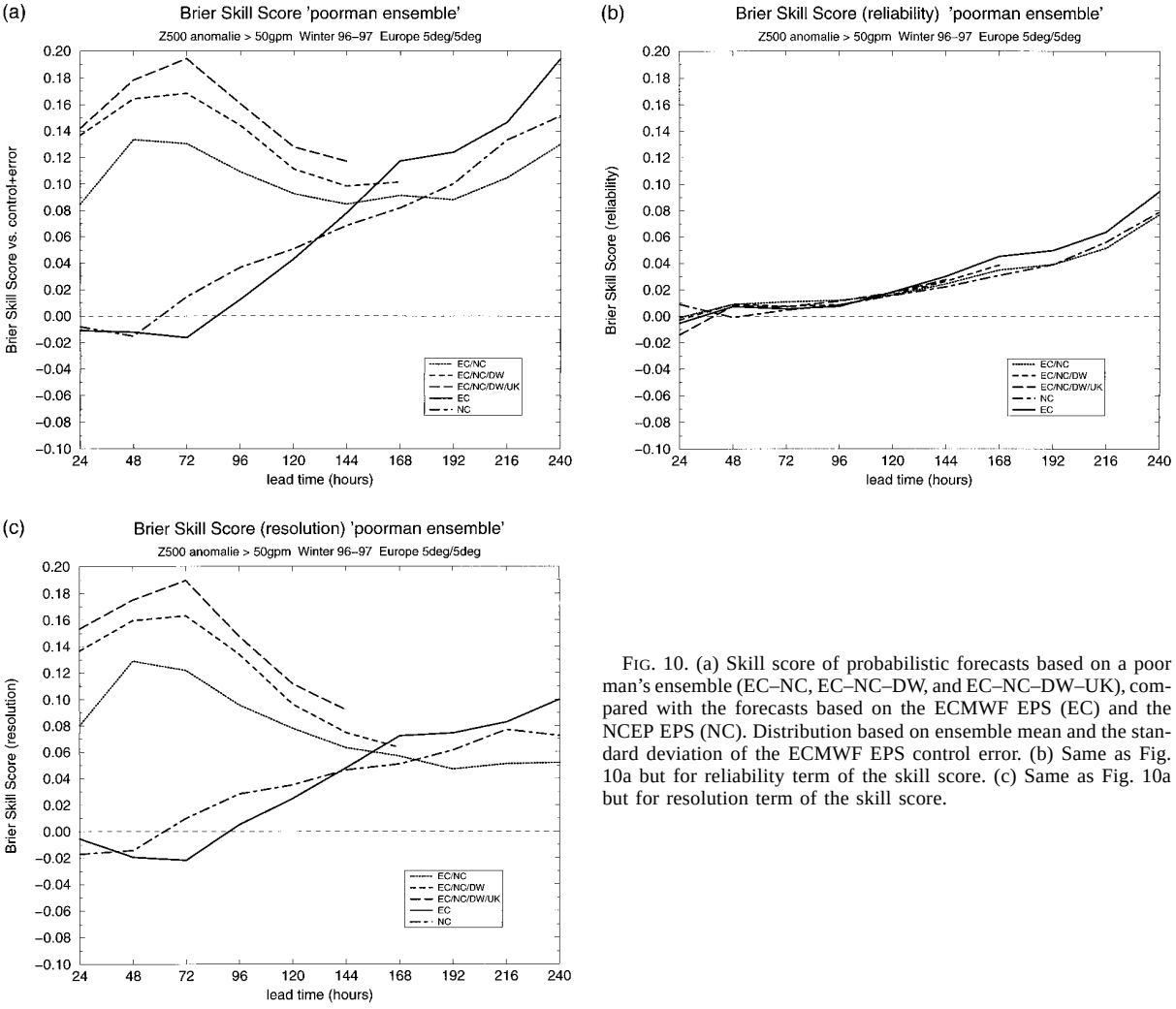


FIG. 10. (a) Skill score of probabilistic forecasts based on a poor man's ensemble (EC–NC, EC–NC–DW, and EC–NC–DW–UK), compared with the forecasts based on the ECMWF EPS (EC) and the NCEP EPS (NC). Distribution based on ensemble mean and the standard deviation of the ECMWF EPS control error. (b) Same as Fig. 10a but for reliability term of the skill score. (c) Same as Fig. 10a but for resolution term of the skill score.

might explain the better performance of ECMWF EPS. One would expect in this case a similar performance of NCEP EPS and ECMWF EPS in winter 1995/96, when the systems were based on models that compared better in terms of resolution. This is not observed (Zhu et al. 1996). Further investigation is needed to fully understand the differences between the performance of the two systems and to determine whether they are significant.

The comparison of the curves labeled EM+error and EM+stdev shown in Fig. 9a clearly indicates a strong negative impact of the spread on the overall performance of the NCEP EPS, even worse than noticed for the ECMWF EPS. This is obviously a consequence of the lack of spread of the NCEP EPS (Toth and Kalnay 1997). This lack of spread is not only due to the small ensemble size. Although Figs. 9a (NCEP) and 6a (ECMWF 10 + 1 members) compare rather well, Table 1 shows that the NCEP EPS spread is on average smaller than the spread of 11 members extracted from the ECMWF ensemble. The difference between the two systems in this respect is probably related to the better model climatology of the ECMWF EPS, partly due to a higher horizontal resolution.

It has been mentioned in previous sections that a lack of spread has little impact on the resolution. Figures 9b and 9c (cf. Figs. 4b and 4c) show that the performance of the two systems is more similar in terms of resolution than in terms of reliability. The better reliability of ECMWF EPS beyond +48 h is obviously a consequence of the spread being larger. Again, this is not only due to a larger number of members, as shown by the comparison of Figs. 6b (ECMWF 10 + 1 members) and 9b (NCEP). On the other hand, the larger spread of NCEP EPS before and until +48 h (due to the breeding method) does not lead to a better reliability.

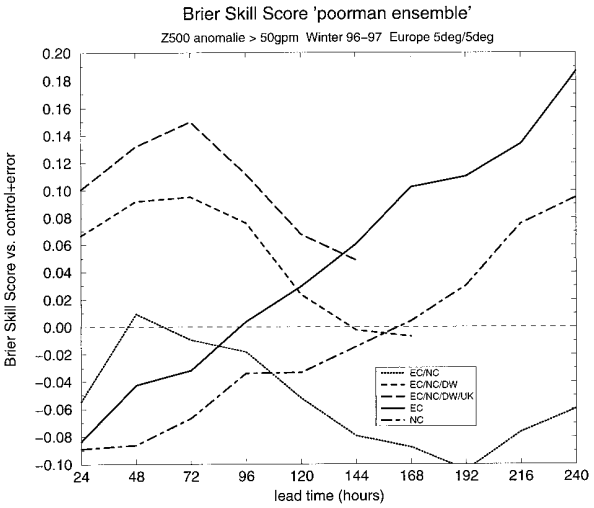


FIG. 11. Same as Fig. 10a but for distribution based on ensemble mean and ensemble standard deviation.

FIG. 12. Same as Fig. 10a but for 150-m geopotential height anomaly.

7. Poor man's ensemble prediction system

In this section the cost efficiency of the ECMWF EPS is addressed by comparison with the performance of a poor man's scheme using current forecasts from various centers, which can be considered as a gratis ensemble prediction system. The poor man's EPS is defined here as the following collection of forecasts: the ECMWF EPS control forecast (EC), the Deutscher Wetterdienst global model forecast (DW), the United Kingdom Meteorological Office unified model forecast (UK), and the NCEP EPS control forecast (NC). Both NC and EC are evaluated out to +240 h, DW to +168 h, and UK to +144 h.

With such a small ensemble (two to four members according to the time step) the score of the raw distribution is unlikely to be significant (see section 5). The poor man's skill score has thus been computed for two distributions: (i) a Gaussian distribution based on the ensemble mean and standard deviation and (ii) a Gaussian distribution based on the ensemble mean and the standard deviation of the control forecast error. The reference forecast is the same as in sections 4 and 5, based on the ECMWF EPS control forecast.

The distribution based on the poor man's ensemble mean and the standard deviation of the control forecast error has an impressively high skill score, much better in the short range and early medium range than the forecasts based on the ECMWF EPS or the NCEP EPS (Fig. 10a). The situation is reversed around +144 h as far as the two-member (EC–NC) poor man's ensemble is concerned. The four-member poor man's ensemble (EC–NC–DW–UK) runs only to +144 h, when it is still more skillful than the ECMWF EPS or the NCEP EPS forecasts. The skill difference is only due to a much better resolution of the poor man's scheme (Fig. 10c). The reliability of the poor man's EPS is equivalent to the reliability of either the ECMWF EPS or the NCEP EPS (Fig. 10b).

The distribution based on the poor man's ensemble mean and standard deviation is not as skillful, especially as far as the two-member scheme is concerned (Fig. 11). This indicates that the skill of the poor man's scheme is exclusively due to its ensemble mean, the spread contributing negatively to the skill. This confirms the results presented in section 5 concerning the impact of reducing ensemble size on the spread contribution to the skill score. Still the distribution based on the four-member poor man's ensemble mean and standard deviation is more skillful than the distributions based on the ECMWF EPS or the NCEP EPS mean and standard deviation up to +144 h.

TABLE 2. Occurrences of 500-hPa geopotential height anomaly exceeding various thresholds, and corresponding skill score and resolution component of the skill score of probabilistic forecasts based on 1) the ECMWF EPS raw distribution and 2) a Gaussian distribution based on the ECMWF EPS mean and the standard deviation of the control forecast error, winter 1996/97, +144 h.

ECMWF EPS (+144 h)	25 m	50 m	75 m	100 m	125 m	150 m
Occurrences	26%	21%	17%	13%	10%	8%
EPS skill	0.06	0.06	0.07	0.08	0.08	0.09
EM + error skill	0.08	0.08	0.08	0.08	0.08	0.08
EPS resolution	0.05	0.05	0.05	0.05	0.05	0.05
EM + error resolution	0.05	0.05	0.04	0.04	0.03	0.03

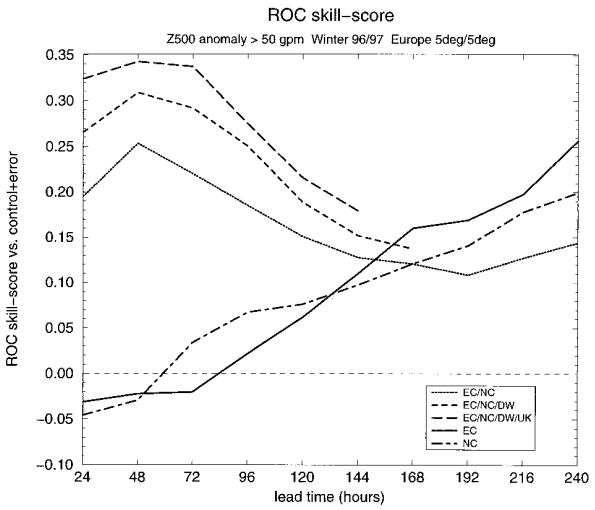


FIG. 13. Same as Fig. 10a but for skill score based on the area below the ROC curve instead of the probability score.

The results presented in this section suggest that it would be worth considering a multianalysis and/or multimodel approach besides operational ensemble prediction techniques used at ECMWF and NCEP, as proposed by Harrison et al. (1995). The benefit of combining numerical forecasts from different NWP centers was demonstrated by Rousseau and Chapelet more than 10 years ago (1985). Ziehmann (1998) presented some evidence that in some circumstances one might get better results using a limited number of forecasts from different NWP centers rather than the ECMWF EPS. Comparable results have been obtained by Talagrand et al. (1998) with a poor man's scheme rather different from the one used in the present study.

8. Summary and concluding remarks

The performance of ensemble prediction systems has been investigated with respect to the probability distribution of 500-hPa geopotential height over Europe. The probability score has been used to assess the performance of probabilistic forecasts. A skill score has been defined by comparison to a reference forecast based on the control forecast of the system under consideration. A decomposition of the skill score has been applied in order to distinguish between the reliability and the resolution aspects of the performance.

The following results have been highlighted:

- The ECMWF EPS forecast is skillful compared to the reference forecast from +96 h onward.
- Most of the ECMWF EPS skill is obtained from a Gaussian distribution based on the ensemble mean. The ensemble spread contributes negatively to the skill score.
- The reliability increases significantly after calibration of the EPS standard deviation, leading to a skillful forecast from +72 h onward.
- The skill score of the ECMWF EPS decreases when reducing ensemble size. This decrease comes mainly from the contribution of the ensemble spread to the score.
- The NCEP EPS has a lower skill score than the ECMWF EPS at all time steps, especially in terms of reliability. This seems a consequence of the insufficient spread of the former.
- The NCEP and ECMWF ensemble means contribute similarly to the skill. The performance of the two systems in terms of resolution is comparable.
- A poor man's EPS is more skillful than either the ECMWF EPS or the NCEP EPS, up to +144 h. This is due to a better resolution, despite a negative contribution of the spread.

These results have been obtained with regard to a single binary event: the 500-hPa geopotential height anomaly exceeding 50 m in magnitude. This threshold has been chosen so that the occurrence of the event is roughly 20% (Table 2). One could argue that a higher threshold would have been a better choice to assess the performance of ensemble prediction systems in providing information on most extreme events. Table 2 shows the ECMWF EPS skill score increasing slightly with a threshold varying from 25 to 150 m, while the resolution component is rather constant. The Gaussian distribution based on the ensemble mean and the standard deviation of the control error is more skillful than the raw distribution for lower thresholds, up to 100 m for the skill score, up to 50 m only for the resolution component. Beyond these limits the contribution of ensemble spread to the skill score becomes positive.

One important issue is the validity of the results presented in section 7 (poor man's EPS leading ECMWF EPS and NCEP EPS) for higher thresholds. Figure 12 shows the same comparison as Fig. 10a for a 150-m threshold, corresponding to a rather *extreme* event since it occurs in less than 8% of occasions (Table 2). The result is the same as pointed out in section 7: the poor man's scheme is better than the ECMWF EPS and the NCEP EPS up to +144 h.

A second limitation of the study is the use of a unique performance criterion, the skill score based on the probability score. Similar conclusions can be drawn from a different verification approach based on the relative operating characteristics (ROC) curve (Mason 1982). The ROC curve is a plot of the hit rate as a function of the false alarm rate of a series of deterministic forecasts, obtained from the probability distribution by considering several probability thresholds, from $p = 0\%$ (event systematically forecast) to $p = 100\%$ (event never forecast). As pointed out by Stanski et al. (1989) the ROC curve says nothing about reliability since it is based on a stratification by observations. Therefore it is not surprising to find that results from this approach are similar

to those concerning the resolution component of the skill score used in this paper. An example of this similarity is shown in Fig. 13, equivalent to Fig. 10 with respect to the area below the ROC curve.

Acknowledgments. Zoltan Toth, Roberto Buizza, Tim Palmer, and François Lalaurette, as well as three anonymous reviewers, provided helpful comments on earlier versions of this manuscript. Acknowledgment is also made to David Stephenson who contributed to the correctness of the text. Special thanks are expressed to Olivier Talagrand for his constant help and support.

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

Buizza, R., 1997: Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125,** 99–119.

——, and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126,** 2503–2518.

——, T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124,** 1935–1960.

Courtier, P., C. Freyder, J. F. Geleyn, F. Rabier, and M. Rochas, 1991: The Arpege project at Meteo-France. *Proc. Seminar on Numerical Methods in Atmospheric Models*, Vol. 2, Reading, United Kingdom, ECMWF, 192–231.

Gibson, J. K., P. Kallberg, S. Uppala, A. Hernandez, A. Nomura, and E. Serrano, 1997: ERA description. ECMWF Re-Analysis Project Report Series, Vol. 1, 72 pp.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126,** 711–724.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, R. Buizza, and T. Petroliagis, 1995: Joint ensembles from the UKMO and ECMWF models. *Proc. Seminar on Predictability*, Vol. 2, Reading, United Kingdom, ECMWF, 61–120.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102,** 409–418.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy Meteor. Soc.*, **122,** 73–119.

Murphy, A., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12,** 595–600.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1993: Ensemble prediction. *Proc. Seminar on Validation of Models over Europe*, Vol. 1, Reading, United Kingdom, ECMWF, 21–66.

Rousseau, D., and P. Chapelet, 1985: A test of the Monte-Carlo method using the WMO/CAS Intercomparison Project data. Report of the second session of the CAS working group on short- and medium-range weather prediction research, WMO/TD 91, PSMP Rep. Series 18, 114 pp.

Simmons, A. J., D. M. Burridge, M. Jarraud, C. Girard, and W. Wergen, 1989: The ECMWF medium-range prediction models development of the numerical formulations and the impact of increased resolution. *Meteor. Atmos. Phys.*, **40,** 28–60.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO/WWW Tech. Rep. 8, 114 pp.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Proc. Seminar on Predictability*, Reading, United Kingdom, ECMWF, 1–26.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125,** 3297–3319.

——, Y. Zhu, T. Marchok, S. Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints, *12th Conf. on Numerical Weather Prediction*, Phoenix, AZ, Amer. Meteor. Soc., 286–289.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8,** 379–398.

Wilson, L. J., 1995: Verification of weather element forecasts from an ensemble prediction system. *Proc. Fifth Workshop on Meteorological Operational Systems*, Reading, United Kingdom, ECMWF, 114–126.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.

Ziehmann, C., 1998: Comparison of the ECMWF ensemble with an ensemble consisting of four operational models. *Abstracts, Seventh Int. Meeting on Statistical Climatology*, Whistler, BC, Canada, ECMWF, 147.

Chapitre 2

Valeur économique de prévisions probabilistes de forts cumuls de précipitation sur la France

Autour de l'article : "Verification of intense precipitation forecasts from single models and ensemble prediction systems" (Atger, 2001)

1. Introduction

Parmi les objectifs affichés des systèmes de prévision d'ensemble, la prévision des phénomènes météorologiques extrêmes n'est apparue que tardivement, après plusieurs années d'exploitation opérationnelle. C'est d'abord la quantification de l'incertitude qui était visée, et, dans une moindre mesure, la prévision de scénarios alternatifs (Palmer et al., 1993). La possibilité de fournir une probabilité d'occurrence pour certains phénomènes n'était pas ignorée, mais l'objectif était plutôt d'identifier les caractéristiques météorologiques les plus marquantes : temps sec ou humide, températures supérieures ou inférieures à la normale, etc. Les premiers produits "probabilistes" destinés aux états membres du CEPMMT à partir du début de l'année 1993 reflétaient cette orientation (Molteni et al., 1996). Il fallait avant tout aider les prévisionnistes à fournir des informations fiables pour la prévision du temps à moyenne échéance (c'est la raison d'être du CEPMMT).

En 1996 une série d'épisodes pluvieux de forte intensité toucha l'Europe. Les inondations qui en résultèrent furent particulièrement dévastatrices dans les pays d'Europe centrale, ainsi qu'en Grèce (Petroliagis et al., 1997). A la même époque le service météorologique Américain s'engageait dans un programme de développement incluant la prévision des fortes précipitations (Fritsch et al., 1998). La question de la prévision des épisodes pluvieux intenses devint centrale dans les discussions portant sur la validation de l'EPS du CEPMMT. Parmi les arguments qui conduisirent à l'augmentation simultanée du nombre de membres

de l'EPS et de la résolution du modèle sous-jacent en décembre 1996, on trouvait déjà en filigrane la volonté de mieux échantillonner les faibles probabilités, même si l'effort principal portait sur l'accroissement de résolution permettant de combler la différence de performance avec le modèle dit "opérationnel", c'est à dire l'intégration unique de la version à haute résolution du modèle (Buizza et al., 1997 ; Buizza & Palmer, 1998). Les deux conditions sont en effet requises lorsqu'on affiche l'ambition de prévoir les épisodes de temps sévère : une résolution élevée permet de modéliser de manière réaliste les phénomènes d'échelle réduite, tandis que l'augmentation du nombre de membres permet un échantillonnage dense des queues de distribution et rend ainsi possible la prévision des faibles probabilités d'occurence des phénomènes rares (Richardson, 2000).

Les années 1997-1998 virent aussi l'émergence, ou plutôt la réapparition après presque deux décennies d'oubli de la part de la communauté météorologique, de la notion de valeur économique des prévisions (Richardson, 2000 ; Richardson, 2001). C'est un concept qui a donné lieu à de multiples publications dans les années 1970-80, en particulier sur le thème de l'utilité des prévisions météorologiques dans le contexte d'une exploitation agricole (par exemple : Katz et al., 1982, ainsi que les 11 autres articles publiés dans le même numéro spécial du *Journal of Applied Meteorolo*gy consacré à la "valeur économique et sociale de l'information sur le temps et le climat"). L'utilisation de modèles simples du processus de prise de décision permet de calculer la valeur relative que possède une prévision météorologique pour son utilisateur. Il s'agit d'un diagnostic relatif : les décisions prises à l'aide de la prévision fournie permettent de gagner de l'argent par rapport aux décisions qui auraient été prises si seule une connaissance *a priori*, par exemple climatologique, était disponible.

Si la valeur économique s'est rapidement imposée dans le champ de la validation des ensembles depuis les premiers travaux de Mylne (1999) et Richardson (2000), c'est probablement parce que ce diagnostic permet de mettre en évidence la

supériorité intrinsèque des prévisions probabilistes par rapport aux prévisions déterministes. Une prévision probabiliste permet en effet à l'utilisateur de prendre une décision en fonction de sa sensibilité au risque, et plus précisément de sa tolérance aux fausses alarmes d'une part, à la non détection du phénomène d'autre part (Murphy, 1985). Par exemple, le risque d'orage n'a pas les mêmes conséquences pour un promeneur et pour un alpiniste. Le premier risque seulement de devoir interrompre sa promenade pour se mettre à l'abri, le second risque sa vie. Une prévision probabiliste d'orage permet à chacun de prendre la décision appropriée : le promeneur sortira même quand le risque est modéré ; l'alpiniste attendra que le risque soit très faible. Au contraire, une prévision déterministe s'adresse à tous les utilisateurs indépendamment de leur sensibilité. L'exemple de la prévision d'orage n'est pas choisi par hasard : si les prévisions (déterministes) d'orage sont si peu appréciées par la population, en France en particulier, ce n'est pas tant parce que le phénomène est "mal prévu" mais bien parce que la vocation de service public de Météo-France impose aux prévisionnistes de mentionner le risque d'orage dès lors que la sécurité de certains est menacée. Ceux dont la sensibilité est plus faible s'en plaignent amèrement, leur tolérance aux fausses alarmes étant largement dépassée pendant chaque saison estivale.

Il est important de noter que c'est généralement la valeur "potentielle" qu'on utilise pour la validation d'un ensemble (Richardson, 2000). C'est le cas dans l'étude présentée ici. Ce qui rend la valeur d'une prévision probabiliste supérieure à celle d'une prévision déterministe, c'est le fait qu'il existe un seuil de probabilité qui maximise la valeur économique pour un utilisateur donné. En théorie ce seuil de probabilité est égal au rapport C/L, c'est le seuil au-delà duquel l'utilisateur ne peut plus prendre le risque de ne pas être protégé. Cette égalité n'est cependant assurée que dans le cas d'une parfaite fiabilité du système (cf. chapitre 1). En pratique, le seuil optimal diffère plus ou moins du rapport C/L suivant le degré de fiabilité. On fait l'hypothèse, dans la plupart des études portant sur la valeur économique, que le système peut être calibré de manière

optimale, c'est à dire que le seuil de probabilité qui conduit à la valeur optimale peut être connu a priori (cf. chapitres 3 et 4 pour la calibration des prévisions probabilistes). Cette hypothèse permet de considérer isolément la composante de la performance qui est liée à la résolution du système (au sens de la décomposition du score de Brier), indépendamment de la composante liée à la fiabilité (cf. chapitre 1).

Dans le modèle de décision le plus simple, celui qui est adopté dans ce chapitre comme dans la plupart des études récentes (c'était moins souvent le cas dans les années 1970-80), la valeur économique est calculée en fonction du rapport coût/perte de l'utilisateur. Ce rapport C/L (cost/loss) caractérise la sensibilité de l'utilisateur : C est le coût de la protection qui est mise en oeuvre pour éviter la perte L. Par définition la valeur économique est maximale pour un rapport C/L égal à la fréquence d'occurrence du phénomène considéré. La performance d'une prévision, évaluée par les indicateurs traditionnels tels que le score de Brier (cf. chapitre 1), augmente généralement avec la fréquence d'occurrence du phénomène considéré : plus le phénomène est rare, plus il est difficile à prévoir. Au contraire, on constate souvent (mais pas toujours) que la valeur maximale augmente avec la rareté du phénomène. En d'autres termes, la valeur augmente pour la prévision d'un phénomène de plus en plus rare, mais seulement pour les utilisateurs dont le rapport C/L de plus en plus petit indique qu'ils sont prêts à accepter une fréquence de plus en plus élevée de fausses alarmes. Il s'agit cependant d'une valeur *relative*, le plus souvent par rapport à une décision prise à partir d'une connaissance climatologique a priori : plus le phénomène est rare, plus il est facile de prévoir une probabilité qui s'écarte valablement de la fréquence climatologique.

Le risque qu'un phénomène extrême se produise a toujours été au centre des préoccupations des météorologistes. C'est même l'anticipation de ces phénomènes qui a justifié, au XIX^{ème} siècle, la création des futurs services publics de prévision météorologique. C'est pour se prémunir de tempêtes telles que celle qui détruisit

la flotte française en Mer Noire en novembre 1854, pendant la guerre de Crimée, que l'Empereur confia à l'astronome Urbain Le Verrier la mission de mettre en place le "Service météorologique international", réseau qui préfigurait l'actuelle organisation de la météorologie opérationnelle dans le monde (Fierro, 1991).

Les prévisionnistes n'ont pas attendu les ensembles opérationnels pour estimer le risque de phénomène extrême. L'opinion d'un prévisionniste, avant qu'elle ne donne lieu à la formulation d'une prévision, est intrinsèquement probabiliste (Murphy, 1993). Les phénomènes extrêmes ne se distinguent qu'en ce que l'enjeu de leur prévision devient élevé quand le risque augmente, rendant la formulation déterministe périlleuse. Au cours de l'élaboration d'une prévision de cumul de précipitations tout prévisionniste est amené à se forger une opinion sur le risque de dépassement de différents seuils, y compris des seuils très élevés qui ne seront probablement pas atteints. La prévision finale, généralement déterministe, ne traduit qu'une petite partie de ce que le prévisionniste a retiré de son analyse de la situation météorologique et de son évolution prévue. En l'absence d'outils spécifiques permettant d'évaluer l'incertitude de la prévision, tels que les ensembles, cette analyse s'appuie sur une connaissance implicite des caractéristiques et de la performance des modèles de prévision déterministe opérationnels. Les biais de prévision, tout d'abord, sont généralement connus par les prévisionnistes, même s'ils ne sont pas quantifiés précisément. Cette connaissance implicite permet par exemple d'estimer le risque d'un cumul de précipitations de 20 mm lorsque la prévision du modèle indique 12 mm. Les limitations du modèles quant à la représentation de la topographie et de l'orographie, en raison d'une résolution toujours trop faible, sont une autre source d'erreur utilisée par les prévisionnistes pour évaluer l'incertitude de la prévision. C'est ainsi l'ensemble du champ de précipitations prévues sur une zone limitrophe qui est utilisé pour estimer un cumul de précipitation en un point, et non la valeur prévue par le modèle en ce seul point (Atger, 1994).

A partir de ce constat qu'un prévisionniste sait estimer le risque de fort cumul de précipitations à partir de données strictement déterministes, une méthode est proposée dans ce chapitre pour évaluer la performance d'un ensemble par rapport à l'intégration unique d'un modèle. La comparaison directe entre prévision probabiliste et prévision déterministe étant fondamentalement inéquitable (voir ci-dessus) il s'agit de construire des prévisions probabilistes comparables à partir de l'ensemble d'une part, d'une prévision déterministe d'autre part. Un des objectifs est de comparer l'apport d'un accroissement de la résolution du modèle et celui d'une augmentation du nombre de membres de l'ensemble. On vise aussi à une estimation plus rigoureuse de la performance "nette" d'un ensemble, indépendamment de la qualité du modèle sous-jacent, c'est à dire l'apport de la stratégie ensembliste retenue. De même qu'on construit des tables de contingence pour plusieurs seuils de probabilité, on propose de construire des tables de contingence pour plusieurs seuils de cumul prévu, pour un cumul observé donné, ainsi que des tables de contingence pour plusieurs seuils de distance entre le point de prévision et le point d'observation. La première technique a été classiquement utilisée pour la validation de modèles de prévision saisonnière (Mason & Graham, 1999), modèles dont les biais de prévision sont généralement importants et mal maîtrisés. La seconde technique est plus originale, elle renvoie aux méthodes d'aggrégation de l'information spatiale qui sont utilisées en prévision par adaptation statistique.

Un aspect important du travail présenté dans ce chapitre est l'utilisation d'un test statistique non paramétrique. S'agissant de comparer la performance de différents systèmes pour la prévision de phénomènes rares, à partir d'un échantillon de données malheureusement très limité en taille, la question de la significativité des résultats s'est rapidement posée au cours de l'étude. Les tests statistiques classiquement utilisés pour la validation de ce type de résultats imposent des hypothèses discutables. La démarche proposée par Hamill (1999) consiste à tirer profit des moyens de calcul aujourd'hui disponibles pour construire la distribution empirique des différences de performance qu'on mesure

entre deux systèmes de prévision de qualité identique. On échantillonne par ce moyen la distribution de probabilité des différences non significatives. La méthode est puissante, polyvalente (elle est ré-utilisée dans le chapitre 3). Elle impose une hypothèse d'indépendance qui est remplie en considérant séparément les précipitations matinales et les précipitations de l'après-midi, distinction qui s'est avérée enrichissante sur le plan de l'interprétation des résultats.

Les systèmes considérés pour l'évaluation sont les ensembles opérationnels du CEPMMT et du NCEP, ainsi que les prévisions déterministes issues des modèles qui leur sont sous-jacents, à la même résolution ou à résolution supérieure. Comme au chapitre 1 on s'intéresse également aux conséquences de la réduction du nombre de membres de l'ensemble, ainsi qu'à la performance d'un ensemble multi-modèles. Les différences entre précipitations du matin et de l'après-midi sont discutées, ainsi que les effets manifestes du sous-échantillonnage.

2. Méthode (section 2)

Cette section est très développée, elle représente près de la moitié de l'article. Après une description des données et du cadre de l'expérience, on introduit successivement les différents concepts utilisés pour l'évaluation. A partir d'une simple table de contingence 2x2, vue comme une représentation de la distribution conjointe des prévisions et observations dans le cas d'une prévision déterministe, on définit le taux de fausse alarme et le taux de détection qui permettent de construire un point sur un diagramme ROC.

Le rapport C/L est ensuite introduit à partir de considérations sur la sensibilité des utilisateurs aux différents aspects de la performance d'une prévision. Suivant le rapport C/L qui caractérise leur sensibilité on montre que les utilisateurs se "positionnent" différemment sur une courbe ROC. Cette analyse est formalisée grâce à la notion de valeur économique qui est introduite par le biais de la dépense moyenne occasionnée par les pertes consécutives aux dommages d'une part, par les coûts de protection d'autre part. L'expression de la valeur relative en

fonction du rapport C/L et des taux de détection et de fausse alarme permet de faire le lien entre les différentes notions présentées.

L'approche multi-seuils est ensuite décrite. A partir de considérations sur le caractère multi-dimensionnel des prévisions météorologiques, on introduit le concept de tables de contingence spatiales, c'est à dire qui font intervenir, pour la détection d'un cumul de précipitations en un point donné, la prévision du cumul de précipitations dans les environs de ce point.

Enfin, la méthode de ré-échantillonnage utilisée pour tester la validité statistique des résultats est décrite en détail (voir en annexe).

3. Principaux résultats (section 3)

3.1. Comparaison ensemble - contrôle (3.1)

Le principal résultat est que l'ensemble du CEPMMT est plus performant que son contrôle à résolution identique (Fig. 9). C'est vrai surtout pour les rapports C/L inférieurs à la fréquence de dépassement du seuil considéré, c'est à dire pour les utilisateurs très sensibles, tolérants les fausses alarmes. C'est pour le seuil 20 mm que la différence est la plus significative. En revanche, en raison du faible nombre de cas considérés, l'avantage de l'EPS sur le contrôle n'est pas statistiquement significatif pour le seuil 50 mm.

Les résultats obtenus pour l'ensemble du NCEP sont décevants (Fig. 10). L'EPS de 12 UTC est moins performant que son contrôle, mais cela n'a rien d'étonnant puisque les 4 membres perturbés sont basés sur un modèle à résolution dégradée (T62) par rapport au contrôle (T126). Ce qui est plus étonnant, c'est que l'ensemble de 00 UTC, dont tous les membres sont basés sur le même modèle à résolution dégradée (T62), n'est pas plus performant que son contrôle. Ce résultat est en contradiction avec ceux obtenus par Toth et al. (1998) ou Zhu et al. (2001) avec un paramètre d'altitude. Il semble indiquer une faiblesse de l'EPS du NCEP pour ce qui concerne la prévision des fortes précipitations. La faible résolution du modèle pourrait être à l'origine de cette contre-performance. L'absence de toute

aptitude à prévoir les fortes précipitations pourrait en effet neutraliser complètement l'impact positif d'une intégration multiple à partir d'états initiaux perturbés.

La comparaison entre l'EPS du CEPMMT et la prévision unique issue du modèle du CEPMMT à haute résolution (T319), qui atteint une performance équivalente, confirme l'importance de la résolution du modèle pour la prévision des fortes précipitations (Fig. 11).

3.2. Comparaison ensemble - ensemble (3.2)

La comparaison entre les deux ensembles opérationnels (à 12 UTC) est à l'avantage de l'EPS du CEPMMT (Fig. 12). Une comparaison de l'EPS du NCEP avec un ensemble de 5 membres extraits de l'EPS du CEPMMT donne des résultats similaires. Ce dernier résultat indique clairement que la forte résolution (T159 contre T62) importe beaucoup plus que le nombre de membres (51 contre 5) lorsqu'il s'agit de prévoir des fortes précipitations. Il est vrai que la méthode mise en oeuvre dans cette étude fait l'hypothèse qu'un prévisionniste évalue l'incertitude de prévision à partir d'autres sources que la seule dispersion de l'ensemble. Les ensembles très peuplés ne sont donc pas systématiquement avantagés.

L'impact d'une limitation du nombre de membres est faible. Avec 21 membres extraits de l'EPS du CEPMMT on atteint pratiquement le même niveau de performance qu'avec l'ensemble complet (Fig. 13). Il faut "tomber" à 11 membres pour que des différences significatives apparaissent. Ce résultat confirme l'importance limitée de la population d'un ensemble (cf. chapitre 1). A nouveau il faut mentionner que ce résultat est indissociable de la méthode de validation qui valorise d'autres sources d'information sur l'incertitude de prévision que la seule dispersion de l'ensemble.

L'EPS du CEPMMT est finalement comparé à un ensemble bi-modèle formé à partir des prévisions de contrôle du CEPMMT et du NCEP. Les résultats sont

partagés. Lorsque les différences sont significatives elles indiquent un léger avantage de l'un ou l'autre des deux ensembles (Fig. 14). Un résultat similaire, pour la même échéance (+96h) mais portant sur la prévision d'une anomalie modérée de géopotentiel en altitude, est présenté au chapitre 1.

4. Discussion (section 4)

Les résultats présentés dans la section précedente concernent les précipitations se produisant l'après-midi (entre 12 UTC et 00 UTC). Les résultats concernant les précipitations matinales sont sensiblement différents (Fig. 15). Les ensembles du CEPMMT et du NCEP sont en effet plus largement avantagés par rapport à leurs contrôles et par rapport à l'ensemble bi-modèle. L'explication proposée est que les précipitations matinales sont davantage associées à des systèmes perturbés de grande échelle, tandis que l'activité convective est surtout présente l'après-midi. Ce déséquilibre est d'autant plus marqué que les données utilisées portent sur une saison d'hiver.

Une des limitations majeures de l'étude est le fait qu'on considère la valeur économique *potentielle* des prévisions. Comme indiqué dans l'introduction à ce chapitre, on fait ainsi l'hypothèse que l'utilisateur connaît a priori le comportement du système de prévision. Lorsque la prévision probabiliste est construite directement à partir du nombre de membres prévoyant l'occurrence du phénomène, on suppose que les sous-estimations ou surestimations des probabilités sont connues. Dans le cas de l'étude présentée ici, on suppose connue la configuration "n membres de l'ensemble prévoyant un dépassement du seuil s à une distance d du point considéré" qui conduit à la valeur économique maximale. Le nombre de configurations possibles étant très élevé, et l'échantillon de données relativement petit, il existe un fort risque de "sur-adaptation" ('over-fitting'), c'est à dire que la valeur économique ainsi calculée ne soit pas seulement potentielle mais aussi complètement virtuelle.

Pour évaluer l'importance de cet effet une évaluation a été menée en divisant l'échantillon disponible en deux sous-échantillons de taille équivalente. On

considère alors que l'utilisateur "apprend" la meilleure configuration à partir du premier échantillon et "l'applique" au second échantillon. Comme les sous-échantillons sont petits (45 jours) les résultats obtenus sont rarement significatifs (Fig. 16). Quand ils le sont ils indiquent plutôt une supériorité relative du contrôle par rapport à l'ensemble. La conclusion de cette discussion est qu'un ensemble possède la capacité de battre son contrôle, mais des échantillons très grands sont nécessaires pour mettre en évidence cette supériorité.

L'importance de disposer de grands échantillons pour évaluer la performance d'un ensemble, en particulier lorsqu'on s'intéresse à des événements rares, est discutée longuement dans les chapitres 3 et 4 de la thèse.

La discussion se termine par un commentaire sur la signification des très petits rapports C/L, et le risque qu'il y aurait à perdre le sens des réalités lorsqu'on évalue la qualité de systèmes de prévision dont la vocation est, avant tout, de permettre à des prévisionnistes d'alerter la collectivité quand il existe un risque météorologique significatif (cf. l'introduction à ce chapitre).

5. Remarques complémentaires

La discussion finale met en évidence la faible significativité des résultats obtenus. La méthode d'évaluation est particulièrement sensible au sous-échantillonnage du fait du grand nombre de membres de l'ensemble (51) et des multiples classes de cumul de précipitation et de distance au point de vérification. L'intérêt de l'étude réside davantage dans l'approche proposée que dans les résultats obtenus. Un prolongement possible de ce travail consisterait à utiliser un échantillon plus large, et surtout à limiter le nombre de degrés de liberté du système considéré, par exemple en réduisant le nombre de catégories de probabilité prévue (problème abordé de manière plus générale dans le chapitre 4 de cette thèse). Il serait également judicieux de distinguer, en terme de performance, l'apport de l'approche spatiale et celle de l'approche multi-seuils.

Annexe

Description de la méthode de ré-échantillonnage utilisée pour tester la validité statistique des résultats obtenus

En matière d'évaluation comparative, une question qui se pose très souvent est la suivante : les différences de score entre deux systèmes de prévision A et B sont-elles significatives, ou bien ces différences sont-elles la conséquence des hasards de l'échantillonnage ? Les événements météorologiques considérés étant peu fréquents, l'échantillon de taille réduite, et le nombre de catégories de prévision très élevé (pour construire les tables de contingence), cette question prend une importance particulière dans la présente étude.

Le principe de la méthode proposée ici est simple : il s'agit d'effectuer un regroupement des deux séries de prévisions issues des systèmes A et B et d'en extraire, par tirage aléatoire, deux séries de prévisions qu'on considère issues des systèmes virtuels A' et B'. La différence de score entre les systèmes A' et B' n'a aucune raison, autre que le hasard de l'échantillonnage, d'être différente de 0. En effectuant un grand nombre de fois cette opération de tirage aléatoire on construit une distribution empirique des différences non significatives entre les scores obtenus par 2 systèmes de prévision. En comparant cette distribution à la différence de score entre les systèmes A et B on déduit une estimation de la probabilité que cette différence soit significative.

Ce test statistique (proposé initialement par Hamill, 1999) est non paramétrique et ne requiert donc d'autre hypothèse que l'indépendance entre les données constituant l'échantillon. Pour éviter les corrélations spatiales on a considéré ensemble toutes les données locales pour une même échéance. Pour réduire les corrélations temporelles on a séparé les précipitations du matin (de 00 à 12 UTC) et les précipitations de l'après-midi (de 12 à 00 UTC).

Les étapes de la procédure sont les suivantes :

1. On construit les tables de contingence des systèmes A et B (par exemple, ECL et ECEPS) pour chacune des N périodes de 12 heures pour lesquelles on dispose de prévisions et d'observations (matin et après-midi séparément). Pour chaque rapport C/L, la valeur relative du système A (resp. B) est calculée à partir des N tables correspondantes. On calcule la différence absolue entre les valeurs relatives des systèmes A et B (pour chaque rapport C/L).

2. En regroupant les tables de contingence construites à l'étape 1 pour les systèmes A et B on obtient un ensemble de 2N tables de contingence (pour chaque rapport C/L).

3. On extrait par tirage aléatoire N tables de contingence de l'ensemble obtenu à l'étape 2, à partir desquelles on calcule la valeur relative d'un système virtuel A' (pour chaque rapport C/L). La valeur relative du système virtuel B' est calculée à partir des N tables de contingence restantes. On calcule la différence absolue entre les valeurs relatives des systèmes A' et B' (pour chaque rapport C/L).

4. L'étape 3 est itérée 1000 fois. On obtient une distribution empirique des différences absolues entre les valeurs relatives de 2 systèmes A' et B' de qualité équivalente.

5. La probabilité que la différence absolue entre les valeurs relatives des systèmes A et B (calculée à l'étape 1) soit significative est estimée à partir de la distribution obtenue à l'étape 4.

Bibliographie

Atger, F., 1994: About quantitative precipitation prediction. Atmospheric physics and dynamics in the analysis and prognosis of precipitation fields, Rome, Italy, AGI/SIMA (Proceedings, 276-280).

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes in Geophysics*, **8**, 401-417.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, N. Wedi, 1997: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935-1960.

Buizza, R. and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503-2518.

Fierro, A., 1991: Histoire de la météorologie. Denoël, Paris, 315 pp.

Fritsch, J.M., R.A. Houze Jr, R. Adler, H. Bluestein, L. Bosart, J. Brown, F. Carr, C. Davis, R.H. Johnson, N. Junker, Y.-H. Kuo, S. Rutledge, J. Smith, Z. Toth, J.W. Wilson, E. Zipser, and D. Zrnic, 1998: Quantitative precipitation forecasting: report of the eighth prospectus development team, U.S. Weather Reasearch Program. *Bull. Amer. Met. Soc.*, **79**, 285-299.

Hamill, T. M, 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

Katz, R.W., A.H. Murphy, and R.L. Winkler, 1982: Assessing the value of frost forecasts to orchardists: a dynamic decision making approach. *J. Appl. Meteor.*, **21**, 518-531.

Mason, S. J. and N. E. Graham, 1999. Conditional probabilities, Relative Operating Characteristics, and Relative Operating Levels. *Wea. Forecasting*, **14**, 713-725.

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73-119.

Murphy, A. H., 1985. Probabilistic weather forecasting. Probability, Statistics, and decision making in the Atmospheric Sciences, A.H. Murphy and R.W. Katz, Eds., Westview Press, 337-377.

Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

Mylne, K., 1999: The use of forecast value calculations for optimal decision making using probability forecasts. Preprints, 17th conference on weather analysis and forecasting, Denver, CO, Amer. Met. Soc., 235-239.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tribbia, 1993: Ensemble prediction. Seminar on Validation of Models over Europe, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, vol. 1, 21-66).

Petroliagis, T., R. Buizza, A. Lanziger and T. N. Palmer, 1997: Potential use of the ECMWF ensemble prediction system in cases of extreme weather events. *Meteor. Appl.*, **4**, 69-84.

Richardson, D. S, 2000. Skill and economic value of the ECMWF Ensemble Prediction System, *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.

Richardson, D. S., 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473-2489.

Toth, Z., Y. Zhu, T. Marchok, S. Tracton and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints, 12th Conf. on Numerical Weather Prediction, Phoenix, Arizona, Amer. Meteor. Soc., 286-289.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson and K. Mylne, 2001: The economic value of ensemble based weather forecasts. Bull. Amer. Meteorol. Soc., 83, 73-83.

Nonlinear Processes in Geophysics (2001) 8: 401–417

Nonlinear Processes in Geophysics

Verification of intense precipitation forecasts from single models and ensemble prediction systems

F. Atger[1]

[1]Météo-France, Toulouse, France

Received: 22 September 2000 – Revised: 3 May 2001 – Accepted: 29 May 2001

Abstract. The performance of single models and ensemble prediction systems has been investigated with respect to quantitative precipitation forecasts. Evaluation is based on the potential economic value of +72 h/+96 h forecasts. The verification procedure consists of taking into account all precipitation amounts that are predicted in the vicinity of an observation in order to compute spatial, multi-event contingency tables. A probabilistic forecast from an ensemble can thus be compared to a probabilistic forecast from a single model run. The main results are the following: (1) The performance of the forecasts increases with the precipitation threshold. High levels of potential value reflect high hit rates that are obtained at the expense of a high frequency of false alarms. (2) The ECMWF ensemble performs better than a single forecast based on the same model, even when the resolution of the ensemble is lower. This is true for the NCEP ensemble as well, but only for morning precipitations. (3) The ECMWF ensemble performs better than the 5-member NCEP ensemble running at 12:00 UTC, even when the population of the former is reduced to 5 members. (4) The impact of reducing the population of the ECMWF ensemble is rather small. Differences between 51 members and 21 members are hardly significant. (5) A 2-member poorman ensemble consisting of the control forecasts of the ECMWF and the NCEP ensembles performs as well as the ECMWF ensemble for afternoon precipitations.

1 Introduction

An important aspect of the performance of weather prediction systems is their ability to accurately forecast intense precipitation events, i.e. those events whose intensity is sufficiently exceptional to cause public disruption. Floods, for instance, represent an important loss for human communities all around the world. The increase in model resolution is believed to be an important factor for improvement with respect to the forecast of intense precipitations (Buizza et al., 1999). The impact of the resolution is particularly in question when comparing a high resolution single model to an ensemble prediction system (EPS) that is generally based on a lower resolution model (Buizza et al., 1997). On the other hand, it has been mentioned that a large number of ensemble members is required for successful detection of rare events (Buizza and Palmer, 1998). A densely populated ensemble distribution seems indeed more adequate than a single model run to detect those events that are located in the tails of the climate distribution.

A number of studies have been devoted, at least partially, to the comparative performance of EPSs and single models with respect to quantitative precipitation forecasts (QPFs). Richardson (2000) compared the relative economic value of a single model to the ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF) with respect to QPF. Zhu et al. (2001) did similar work in the U. S. for the National Centers for Environmental Prediction (NCEP) operational forecasting system. In these studies, deterministic forecasts based on a single model are compared to probabilistic forecasts based on an EPS. The information content of a probabilistic forecast is essentially higher than that of a deterministic forecast, since it allows the user to select the right probability threshold that corresponds to his concern (Murphy, 1985). The results of most comparative studies are thus not surprising: EPS probabilistic forecasts are more accurate, skillful and valuable than deterministic single model forecasts (Toth et al., 1998).

Operational forecasters, however, use information from a single model as probabilistic guidance. This is particularly obvious when dealing with extreme events, whose a priori probability is very low, such as intense precipitations. Physical processes that are involved in extreme precipitation events are not taken into account very well in atmospheric models, due to approximations introduced, for example, by the parameterization of the convection, the limited horizontal and vertical resolution, and the poor representation of topography. As a consequence, events that rarely occur are

Correspondence to: F. Atger (frederic.atger@meteo.fr)

even less often predicted by operational models. However, the lack of performance of numerical models in that respect has never prevented operational forecasters to successfully forecast rare, extreme events, on occasion. Forecasters are apparently able to extract from the model output information indicating that an extreme event, although not explicitly predicted by the model, might still occur with significant probability.

Forecasters' judgments are essentially probabilistic (Murphy, 1993), even when the technical information available is purely deterministic, as is the case when a forecaster interprets a single model output. In the case of QPF, the forecaster's judgment can take the form of probabilities of certain thresholds being reached. For example, given a model forecasting of 10 mm/12 h, a forecaster might consider a 5% probability of less than 1 mm/12 h and an 80% probability of more than 8 mm/12 h, with the numbers depending on expected model biases and uncertainties. This probabilistic judgment is not necessarily stated explicitly, but it represents the basis of any statement, including the very deterministic "12 mm/12 h" that may be required for operational purposes. Furthermore, an experienced forecaster would not elaborate a QPF at a given location from the precipitation amount predicted by the model at only that location. Forecasters are well aware of the limitations of numerical weather prediction, especially the consequences of insufficient resolution or poor representation of the topography, as well as the effect of errors in the initial conditions. They generally consider the whole model output in order to obtain an opinion about the expected value of a meteorological variable at a given point. In other words, a forecaster takes advantage of the spatial distribution of a forecast variable in order to predict its local probability density function (pdf). In practice, a high amount of precipitation predicted by the model at a short distance from a given point may indicate a considerable risk of a high precipitation amount, even if no precipitation is predicted by the model at that point. The forecaster's experience, as well as considerations about orography, the expected weather pattern, the model resolution and other characteristics, play an important role in the way this indication is inferred from the available information. The forecaster's judgment can still be facilitated by the use of model output statistics (MOS), especially when explicit probabilistic forecasts are required (Carter, 1989).

Probabilistic EPS forecasts perform undoubtedly better at all lead times than deterministic forecasts based on a single model (Zhu et al., 2001). On the other hand, it has been shown that it is possible to beat an ensemble at early lead times with a probabilistic forecast based only on a single model output and model statistics, when considering upper level variables, such as 850 hPa temperature (Talagrand, 1997) or 500 hPa geopotential height (Atger, 1999). In the present article, a forecast procedure is designed in order to mimic the way in which an operational forecaster infers a QPF from a single model output. Single models with different resolutions, operational ensembles and a "poorman ensemble" (consisting of single model runs) are compared in order to: (i) assess the performance of operational forecasting systems for the prediction of intense precipitations; (ii) evaluate the usefulness of an EPS when used in conjunction with one or several higher resolution models; (iii) investigate the relative impact of model resolution and ensemble population on the performance of an EPS.

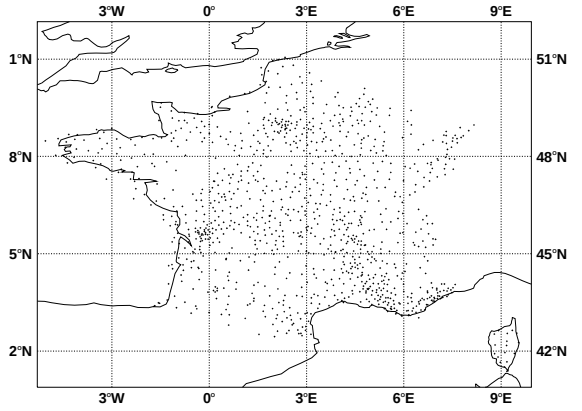


Fig. 1. The French network of rain gauges used in this study. The 1194 stations have reported at least one 12 h-precipitation amount during the winter of 1998–1999.

The article is organized as follows. The methodology is described in Sect. 2. The results are presented in Sect. 3, discussed in Sect. 4, and summarized in Sect. 5.

2 Methodology

2.1 Data

2.1.1 Observations

Observed precipitation data from the French rain gauges network have been collected from winter 1998–1999, i.e. 90 days from 1 December 1998 to 28 February 1999. Original data are 6 h accumulations at 1194 stations in France (Fig. 1). The final set consists of 12 h accumulations from 00:00 UTC to 12:00 UTC, and from 12:00 UTC to 00:00 UTC every day. Selected observations have successfully passed quality controls, so that gross departures from the climate are excluded. Due to missing or rejected data, the final set contains 194 191 values.

2.1.2 Forecasts

The verification procedure has been applied to single runs from the ECMWF model which was operational in winter 1998–1999 (Simmons et al., 1989; Courtier et al., 1991) in its high resolution version ECH (T_L319) and its lower resolution, ensemble prediction version ECL (T_L159). Both versions run at 12:00 UTC. The verification procedure has also been applied to single runs from the NCEP model, running at 12:00 UTC NC12 (T126 resolution up to +84 h,

T62 resolution afterwards) and at 00:00 UTC NC0 (T62 resolution). Concerning ensemble prediction, the verification procedure has been applied to the ECMWF EPS (Palmer et al., 1993; Molteni et al., 1996), which consists of 51 integrations of the T_L159 ECMWF model running at 12:00 UTC (ECEPS) and to the NCEP EPS (Tracton and Kalnay, 1993; Toth and Kalnay, 1997), which consists of 5 integrations of the T62 NCEP model running at 12:00 UTC (NCEPS12) and of 11 integrations of the T62 NCEP model running at 00:00 UTC (NCEPS0). Smaller ensembles have been constructed from the ECMWF EPS by retaining the control forecast and the first 10, 20, 32 perturbed members (ECEPS11, ECEPS21, ECEPS33). A 2-member "poorman ensemble" (ECNC) has been constructed from the single model runs described above. It consists of the ECMWF T159 model forecast ECL and the NCEP T126/T62 model forecast NC12.

Prior to verification, all forecasts have been interpolated onto the same 2.5°/2.5° grid that roughly corresponds to the horizontal resolution of the NCEP T62 model (the lowest resolution of the considered models). Forecasts have been retrieved over a large area surrounding France (56° N, 12° W, 36° N, 15° E) so that forecast data are available in a 500 km circular area around every available observation. In this section, precipitation accumulated from +72 h to +84 h and from +84 h to +96 h have been considered together. Morning precipitations (valid from 00:00 UTC to 12:00 UTC) and afternoon precipitations (valid from 12:00 UTC to 00:00 UTC) have been verified separately in Sect. 3.

2.1.3 Observed and forecast distribution

A cumulative distribution of 12 h precipitations has been computed from the whole set of selected observations. Approximately 100 cases (0.05%) have been identified with an accumulation exceeding 50 mm, 1000 observations (0.5%) with an accumulation exceeding 20 mm, and 10 000 observations (5%) with an accumulation exceeding 5 mm. Since the definition of an intense 12 h precipitation event is rather arbitrary, the 5 mm, 20 mm and 50 mm thresholds have been used as detection thresholds for verification in this study.

For precipitation thresholds from 1 mm to 50 mm (12 h accumulation), Fig. 2 shows the cumulative distribution of the observations and the corresponding forecasts of the different models used in the study, obtained from a bilinear interpolation at the observations. The impact of model horizontal resolution is clearly visible, with the forecast distribution that is closer to the observed distribution corresponding to ECH, the ECMWF T319 model. Note that different results might have been obtained with forecasts interpolated onto a more accurate grid, especially for ECH, whose horizontal resolution is much sharper than the 2.5°/2.5° grid used in the study.

Figure 2 also shows that morning precipitations (00:00–12:00 UTC) and afternoon precipitations (12:00–00:00 UTC) have a different distribution. Intense precipitations are more frequent during the afternoon, probably because convection is more important after 12:00 UTC (this effect would have probably been emphasized if a summer season had been included in the considered period).

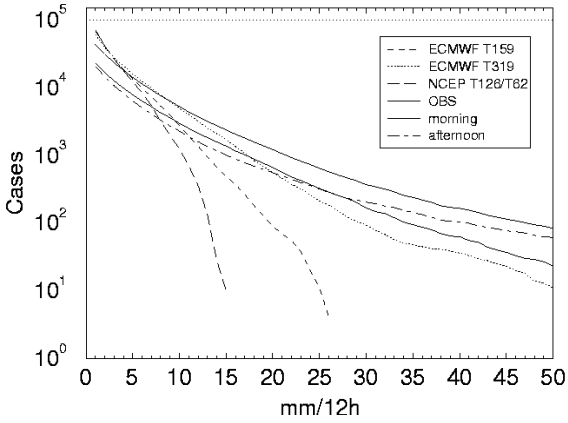


Fig. 2. Cumulative distribution of observed and forecast precipitation amounts for 50 thresholds from 1 mm/12 h to 50 mm/12 h. Observations: all (thick solid line), 00:00–12:00 UTC (thin solid line), 12:00–00:00 UTC (thin dash-dotted line). Forecasts (+84 h) interpolated at all observations: ECMWF T_L159 model (dashed line), ECMWF T_L319 model (dotted line), NCEP T126 model (long dashed line).

2.2 Contingency tables and relative operating curve

In a wide sense, forecast verification consists of a comparison of a distribution of forecasts $p(f)$ to a distribution of observations or analyses $p(x)$. The level of correspondence between $p(f)$ and $p(x)$ indicates how accurate the forecasting system is. There exist a number of methods to estimate this level of correspondence; the most widely used is the computation of the moments of the distribution of errors $p(f-x)$, which leads to the mean error, the mean square error, the standard deviation of the error, etc. The most informative approach, however, consists of a double factorization of the joint distribution of forecasts and observations (Murphy and Winkler, 1987).

$$p(f,x) = p(x|f)p(f) = p(f|x)p(x)\ , \qquad (1)$$

where the joint distribution $p(f,x)$ contains all of the information that is needed to evaluate the forecast's accuracy. By stratifying the data according to the forecasts, the joint distribution can be seen as the product of the distribution of forecasts $p(f)$ and the conditional distribution of observations, given the forecast $p(x|f)$. Similarly, by stratifying the data according to the observations, the joint distribution can be seen as the product of the distribution of observations $p(x)$ and the conditional distribution of forecasts, given the observation $p(f|x)$.

In the case of the deterministic forecast of a meteorological event, e.g. a precipitation amount exceeding 5 mm/12 h, the joint distribution is most generally represented as a 2×2

Table 1. Contingency table based on ECL (ECMWF T159 model, +72 h to +96 h) for the 5 mm/12 h observed threshold. H: number of Hits. FA: number of False Alarms. M: number of Misses. CR: number of Correct Rejections. $HR = H/(H+M) = 0.29$ (Hit Rate). $FAR = FA/(FA+CR) = 0.05$ (False Alarm Rate)

ECL 5 mm/12 h	Observed	Not observed
Forecast	$H = 4094$	$FA = 9426$
Not forecast	$M = 10061$	$CR = 170610$

contingency table. This table indicates, for a given observed (not observed) event, the number of times this event was predicted (non-predicted). Table 1 shows, for example, the contingency table for the 5 mm/12 h threshold and ECL. From this table, the stratification according to observations leads to two useful indicators: the hit rate (HR), which is the proportion of observed events that were successfully predicted and the false alarm rate (FAR), which is the proportion of non-observed events that were erroneously predicted.

In the case of a probabilistic forecast, the joint distribution can be represented similarly as a contingency table built from a number of probability categories. This table indicates, for a given observed (non-observed) event, the number of times every probability category is predicted (non-predicted). When verifying EPS forecasts, the categories are generally defined according to the number of ensemble members that forecast the event, from 1 to N (if N is the number of ensemble members). Table 2 shows, for example, an extract of the contingency table for the 5 mm/12 h threshold and ECEPS for a selection of probability categories based on the number of ensemble members. HR and FAR are computed separately for every category, so that the contingency table leads to an ensemble of pairs (FAR, HR). Every pair indicates the performance of a deterministic forecast that would be based on the fact that at least a certain number of ensemble members forecast the considered event.

It is convenient to plot these (FAR, HR) pairs as an ensemble of points on a diagram, forming the so-called Relative Operating Curve (ROC) (Mason, 1982). The relative position of the ROC obtained from a probabilistic forecasting system and the single point (FAR, HR) obtained from a deterministic forecasting system indicates their relative accuracy (Stanski et al., 1989). A single point above (below) the curve indicates that the deterministic system is more (less) accurate than the probabilistic system. Similarly, the relative position of the ROCs obtained from two probabilistic forecasting systems indicates their relative accuracy. Figure 3 shows, for example, the ROC for the 5 mm threshold and ECEPS. The (FAR, HR) point for ECL is plotted on the same figure. The position of the latter with respect to the former indicates a very similar overall performance of the two systems. Nevertheless, higher HRs (lower FARs) are attained by ECEPS for certain probability categories at the expense of higher FARs (lower HRs). For example,

Table 2. Contingency table based on ECEPS (ECMWF EPS, +72 h to +96h) for the 5 mm/12 h observed threshold. H_j (FA_J): number of Hits (False Alarms) for more than j members forecasting the event. $HR_j = H_j/\Sigma H_j$ (Hit Rate). $FAR_j = FA_j/\Sigma FA_j$ (False Alarm Rate). Example: $HR_2 = 0.78$; $FAR_2 = 0.28$. The number of forecast categories is 51 (ensemble members)

ECEPS 5 mm/12 h	Observed	Not observed
Forecast at least by 1 forecast	$H_1 = 12263$	$FA_1 = 67534$
Forecast at least by 2 forecasts	$H_2 = 11031$	$FA_2$50410
⋮	⋮	⋮
Forecast at least by j forecasts	H_j	FA_j
⋮	⋮	⋮
Forecast at least by 40 forecasts	$H_{40} = 105$	$FA_{40} = 123$
⋮	⋮	⋮
Forecast at least by 51 forecasts	$H_{51} = 0$	$FA_{51} = 0$

$HR = 0.78$ and $FAR = 0.28$ for the second probability category based on ECEPS ("at least 2 members are forecasting more than 5 mm/12 h"), while $HR = 0.29$ and $FAR = 0.05$ for the deterministic forecast based on ECL.

2.3 The cost-loss ratio

Figure 3 shows clearly that the advantage of a probabilistic forecast comes primarily from the fact that certain probability categories lead to higher HRs or lower FARs than those obtained with a single deterministic forecast. This is at the expense of an increase in the FAR or a decrease in the HR. For certain forecast users, a higher HR is valuable enough to tolerate a larger number of false alarms, typically when a user has the power to avoid a high loss L by protecting at low cost C. An example is given by the protection of Bordeaux vineyards from the frost in early spring: given the importance of the potential loss and the relatively low cost of protection, vineyards are protected as soon as the risk of frost exists, even when this risk is low. The so-called cost-loss ratio C/L is low. Another extreme example of low C/L is the protection of human life in the case of the risk of a dangerous meteorological event (e.g. storm, flood). The loss of a human life is incredibly high and the cost to protect it is generally low, so that C/L tends toward zero.

Other users do not tolerate false alarms. Due to a high C/L, they require a FAR as small as possible, even if this condition implies a decrease in the HR. High C/L are typical of long-term decision making situations, for example, the management of energy production: activation/deactivation of a nuclear reactor unit costs a lot, but the expected loss (or benefit, in this case) is limited.

Although all forecast users would benefit from points of the ROC that are ideally located close to the top left corner of Fig. 3, high and low C/L users do not benefit from the same part of the curve; low C/L users benefit from points

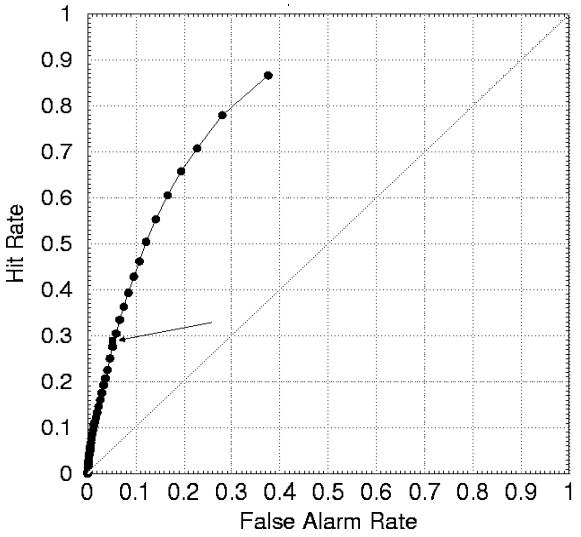


Fig. 3. Relative Operating Curve (ROC) for the 5 mm/12 h observed threshold, drawn from the contingency table shown in Table 2, based on ECEPS (ECMWF EPS, +72 h to +96 h). From the top of the curve, every point indicates the performance of a deterministic forecast based on the fact that at least 1, 2, 3, etc., ensemble members forecast at least 5 mm/12 h (51 points). The single square indicated by the arrow is the (FAR, HR) point drawn from the contingency table shown in Table 1, based on ECL (ECMWF T159 model, +72 h to +96 h). It indicates the performance of a deterministic forecast based on the fact that the model forecasts at least 5 mm/12 h.

of the ROC that are located in the upper part of the curve (higher HR), while high C/L users benefit from points in the lower part of the curve (lower FAR).

2.4 Relative value

Ångstrøm (1919) was probably the first who introduced the concept of value in the field of weather forecasting (Liljas and Murphy 1994). After Murphy (1977), several authors explored multiple aspects of the usefulness/value problem in the 70's and 80's (Katz and Murphy, 1997). According to this approach, users of weather forecasts are "decision makers": they have to take different decisions according to the expected weather conditions. The usefulness of a weather forecast can thus be quantified by considering the occasions when the use of the forecast has been beneficial, detrimental or neutral to the user, with respect to the process of decision making.

Here we consider the particular situation when a user requires a forecast in order to avoid potential damages caused by adverse weather conditions, e.g. intense precipitations. A simple economic model can be applied when the user has just two alternatives: to protect or to do nothing. The cost of protection C is known, as well as the expected loss L occurring in case of damage. If no weather forecast is available, the decision to protect is likely to be based on climate knowledge. If f_c is the expected climatological frequency of the event, it is easy to show that the user should always protect if $C/L < f_c$, otherwise the user should never protect. Let f be the actual frequency of the event during the considered period. Under the assumption that $f_c = f$, the mean expense (per unit loss) $\mathrm{ME}_{\mathrm{climate}}$ is, therefore, the min of C/L and f.

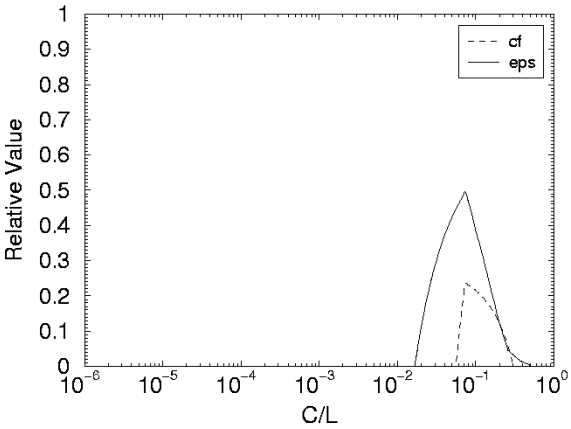


Fig. 4. Relative value for the 5 mm/12 h observed threshold, as a function of the user C/L, based on the contingency tables shown in Table 1 and Table 2. Dash line: ECL (ECMWF T159 model, +72 h to +96 h). Solid line: ECEPS (ECMWF EPS, +72 h to +96 h). By construction, the maximum value is obtained in both cases for $C/L = f$, the frequency of occurence of the event (see Eq. 3).

On the other hand, a perfect knowledge of the future weather would allow the user to protect only when the event occurs, so that $\mathrm{ME}_{\mathrm{perfect}}$ would be the product of C/L and f. The relative economic value of a weather forecast (V) is defined as the amount of money that is saved by the user, normalized by the amount of money that he could save by using a perfect (hypothetical) forecast:

$$V = \frac{\mathrm{ME}_{forecast} - \mathrm{ME}_{climate}}{\mathrm{ME}_{perfect} - \mathrm{ME}_{climate}} \,. \tag{2}$$

Relative value is thus a skill-score based on the mean expected expense, according to the usual definition of the forecast skill (e.g. Stanski et al., 1989). The maximum value $V = 1$ is obtained by a perfect forecast, and $V = 0$ for the climate forecast. From the above discussion about the relative importance of higher HR and FAR for different categories of users, the relative value can be expressed as a function of the user's C/L on the one hand, and as a function of the forecast FAR and HR, on the other hand. Richardson (2000) has demonstrated the following relation:

$$V = \left(\min\left(\frac{C}{L}, f\right) - FAR\frac{C}{L}(1-f) + HR \cdot \left(1 - \frac{C}{L}\right) f - f \right) \left(\min\left(\frac{C}{L}, f\right) - f\frac{C}{L} \right)^{-1} \,. \tag{3}$$

It is important to note that this formulation is correct under the assumption that $f_c = f$, as mentioned above. In prac-

Table 3. Multi-event contingency table based on ECL (ECMWF T159 model, +72 h to +96 h) for the 5 mm/12 h observed threshold. H_k (FA_k): number of Hits (False Alarms) for the k mm/12 h forecast threshold. $HR_k = H_k/\Sigma H_k$ (Hit Rate). $FAR_k = FA_k/\Sigma FA_k$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds)

ECL 5 mm/12 h	Observed	Not observed
Forecast > 1 mm/12 h	$H_1 = 11797$	$FA_1 = 60225$
Forecast > 2 mm/12 h	$H_2 = 9312$	$FA_2 = 34145$
⋮	⋮	⋮
Forecast > k mm/12 h	H_k	FA_k
⋮	⋮	⋮
Forecast > 5 mm/12 h	$H_5 = 4094$	$FA_5 = 9426$
⋮	⋮	⋮
Forecast > 20 mm/12 h	$H_{20} = 5$	$FA_{20} = 70$

Table 4. Multi-event contingency table based on ECEPS (ECMWF EPS, +72 h to +96 h) for the 5 mm/12 h observed threshold. $H_{j,k}$($FA_{j,k}$): number of Hits (False Alarms) for more than j members forecasting more than k mm/12 h. $HR_{j,k} = H_{j,k}/\Sigma H_{j,k}$ (Hit Rate). $FAR_{j,k} = FA_{j,k}/\Sigma FA_{j,k}$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds) × 51 (ensemble members) = 1020

ECEPS 5 mm/12 h		Observed	Not observed
⋮	⋮	⋮	⋮
Forecast > k mm/12 h	At least by 1 forecast	$H_{k,1}$	$FA_{k,1}$
Forecast > $k$mm/12 h	⋮	⋮	⋮
Forecast > k mm/12 h	At least by 51 forecasts	$H_{k,51}$	$FA_{k,51}$
⋮	⋮	⋮	

tice, f is not known before the end of the verification period. The climate forecast is based only on the knowledge of f_c and is not as reliable as it might be if it was based on the knowledge of f, the actual frequency of occurrence of the event. $ME_{climate}$ is, therefore, underestimated in Eq. (2), which has a slight impact on the computed value. The above formulation has, however, been used in most studies, since the computation can be done from the sample only, with no need for independent climatological data. It has been used in the present study for the same reasons.

When considering a probabilistic forecast, there are as many (FAR, HR) pairs as probability categories. For a given C/L, it is, therefore, convenient to consider the maximum value that is attained for the probability category that is optimal for the user, i.e. that leads to the better compromise between a low FAR and a high HR (Richardson, 2000). Figure 4 shows, for example, the value as a function of C/L, for the 5 mm/12 h observed threshold, for the deterministic forecast based on ECL and the probabilistic forecast based on ECEPS (same forecasts as Fig. 3). The ECEPS curve is, in fact, the envelope of the 51 curves of value that are obtained for every forecast category, from "at least 1 member forecasting the event" to "all members forecasting the event". The better performance of the probabilistic forecast based on ECEPS is clearly visible, especially for lower C/L.

2.5 Multi-event contingency tables

In the previous subsection, it has been described how the performance of a deterministic forecast based on a single model can be investigated from a simple 2 × 2 contingency table, which gives a simplified representation of the joint distribution of forecasts and observations, limited to the forecasts and observations of one specified event (see Sect. 2.2). Multi-event contingency tables give a more complete representation of the joint distribution. A table indicates, for a given observed (non-observed) event, the number of times different events are predicted (non-predicted). This approach has been followed in seasonal prediction verification studies (e.g. Mason and Graham, 1999).

Table 3 shows, for example, the multi-event contingency table of ECL for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h. Higher forecast thresholds are not used since they occur very rarely, partly due to the coarse interpolation grid that has been used. Similar to a probabilistic forecast contingency table (e.g. Table 2), a multi-event contingency table leads to several (FAR, HR) pairs, each of which indicates the performance of a deterministic forecast that would be based on the fact that a specified forecast threshold is reached by the model. Therefore, the ensemble of (FAR, HR) pairs indicates the performance of a probabilistic forecast based on a single model run. Figure 5a shows the ROC corresponding to Table 3. The performance is very similar to that shown in Fig. 3, corresponding to the probabilistic forecast based on ECEPS.

Multi-event contingency tables can be used for the verification of probabilistic forecasts based on an EPS as well. A table indicates, for a given observed (non-observed) event, the number of times different events are predicted (non-predicted) by at least a certain number of ensemble members. Table 4 shows, for example, an extract of the multi-event contingency table of ECEPS for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h. Figure 5b shows the ROC corresponding to Table 4. The performance is improved, compared to Fig. 3 (ECEPS) and Fig. 5a (ECL multi-event), in the upper part of the curves where forecasts are primarily beneficial to lower C/L users.

Figure 6 shows the relative value as a function of C/L for the multi-event contingency tables based on ECL and

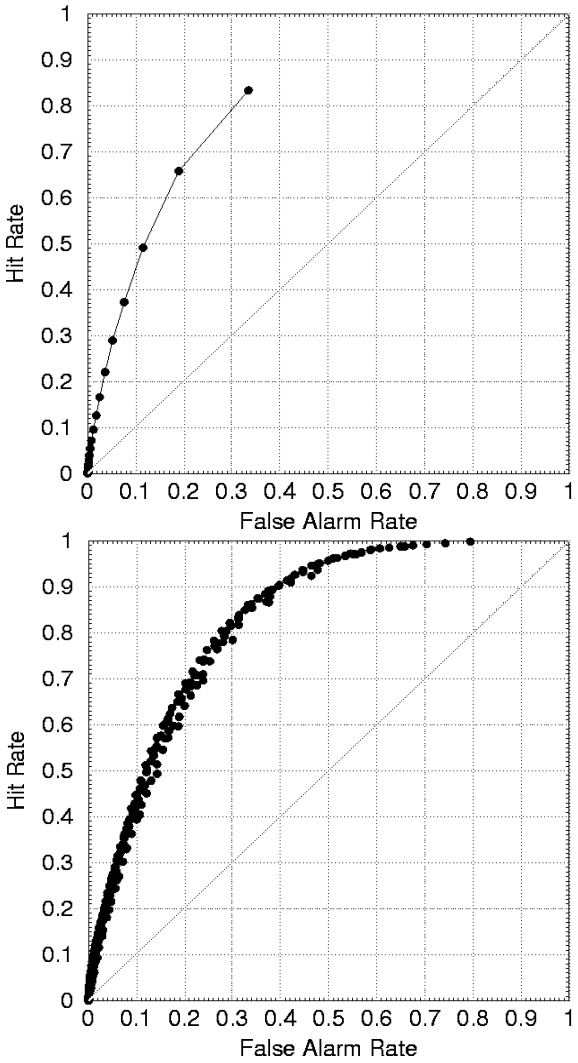


Fig. 5. Relative Operating Curve (ROC) for the 5 mm/12 h observed threshold, based on the multi-event contingency tables shown in Table 3 and Table 4. **(a)** ECL (ECMWF T159 model, +72 h to +96 h); from the top of the curve, every point indicates the performance of a deterministic forecast based on the fact that the model forecasts at least 1, 2, 3, etc., mm/12 h (20 points). **(b)** ECEPS (ECMWF EPS, +72 h to +96 h); every point indicates the performance of a deterministic forecast based on the fact that at least 1, 2, 3, etc., ensemble members forecast at least 1, 2, 3, etc., mm/12 h (51×20=1020 points).

ECEPS (same forecasts as in Fig. 5a and Fig. 5b). The ECL curve is, in fact, the envelope of the 20 curves of value that are obtained for every forecast threshold, from 1 mm/12 h to 20 mm/12 h. The ECEPS curve is the envelope of the 20×51=1020 curves of value that are obtained for every forecast category, from "at least 1 member forecasting more than 1 mm/12 h" to "all members forecasting more than 20 mm/12 h". The forecast based on ECEPS is only slightly

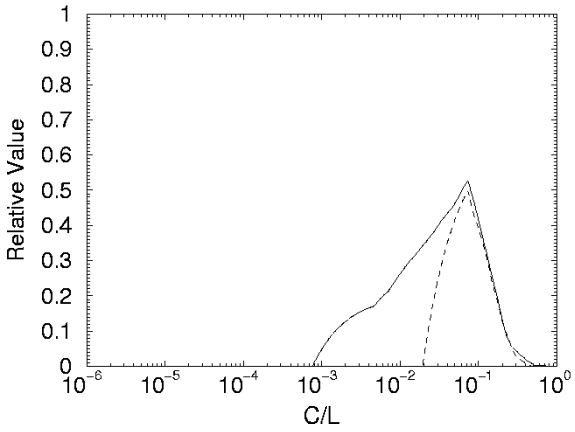


Fig. 6. Relative value for the 5 mm/12 h observed threshold, as a function of the user C/L, based on the multi-event contingency tables shown in Table 3 and Table 4. Dash line: ECL (ECMWF T159 model, +72 h to +96 h). Solid line: ECEPS (ECMWF EPS, +72 h to +96 h).

better overall than the forecast based on ECL, but is much better for lower C/L.

2.6 Spatial contingency tables

Verification of QPF as well as most quantitative weather forecasts would ideally require one to consider the correspondence between 3-dimensional distributions of forecasts and observations: two dimensions for the physical space, and one dimension for time. In practice, verification generally consists of an evaluation of the correspondence between (time distributions of) local forecasts and local observations, as described in the previous subsections. Space connections between local forecasts and local observations are rarely considered. A spatial approach of verification would consist of an evaluation of the correspondence between (time distributions of) spatial distributions of forecasts and spatial distributions of observations. One application of this approach is the evaluation of the correspondence between forecast and observed meteorological patterns, through the computation of Anomaly Correlation or the categorization of large-scale circulation patterns (Chessa and Lalaurette, 2000). Another application is the evaluation of the correspondence between a local observation and the local forecasts that are found in the vicinity of this observation.

Spatial multi-event contingency tables have been used in the present study. Each table indicates, for a given observed (non-observed) event, the number of times different events are predicted (non-predicted) at different distances from the observed event. Table 5 shows, for example, an extract of the spatial multi-event contingency table of ECL for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h, at 100 km, 200 km, 300 km, 400 km and 500 km from the observation.

Table 5. Spatial multi-event contingency table based on ECL (ECMWF T159 model, +72 h to +96 h) for the 5 mm/12 h observed threshold. $H_{k,l}(FA_{k,l})$: number of Hits (False Alarms) for more than k mm/12 h forecast at less than l×100 km from the observation. $HR_{k,l} = H_{k,l}/\Sigma H_{k,l}$ (Hit Rate). $FAR_{k,l} = FA_{k,l}/\Sigma FA_{k,l}$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds) × 5 (distances to the observation) = 100

ECL 5 mm/12 h		Observed	Not observed
⋮	⋮	⋮	⋮
Forecast $> k$ mm/12 h	At less than 100 km from observation	$H_{k,1}$	$FA_{k,1}$
Forecast $> k$ mm/12 h	⋮	⋮	⋮
Forecast $> k$ mm/12 h	At less than 500 km from observation	$H_{k,5}$	$FA_{k,5}$
⋮	⋮	⋮	⋮

Figure 7a shows the ROC corresponding to Table 5. Although most of the points of Fig. 7a are located below those of Fig. 5a, the envelope of the curves is almost identical, except in the upper part of the curve where a higher HR can be obtained at the expense of a higher FAR. This part of the curve is obtained with forecast categories that are very "sensitive" in detecting the occurrence of rain, with the most sensitive at 1 mm/12 h at 500 km from the observation.

Table 6 shows an extract of the spatial multi-event contingency table of ECEPS for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h, at 100 km, 200 km, 300 km, 400 km and 500 km from the observation. Figure 7b shows the ROC corresponding to this contingency table. Again, most of the points of Fig. 7b are located below those of Fig. 5b, but the higher density of points leads to an envelope that is slightly better. Figure 8 shows the relative value as a function of C/L for the spatial multi-event contingency tables based on ECL and ECEPS (same forecasts as in Fig. 7a and Fig. 7b). The curves are, in fact, the envelopes of the curves of value that are obtained for every forecast category (5 × 20=100 categories for ECL, 5 × 20 × 51=5100 categories for ECEPS). The forecast based on ECEPS is better than the forecast based on ECL for lower C/L, but the difference is reduced compared to Fig. 6.

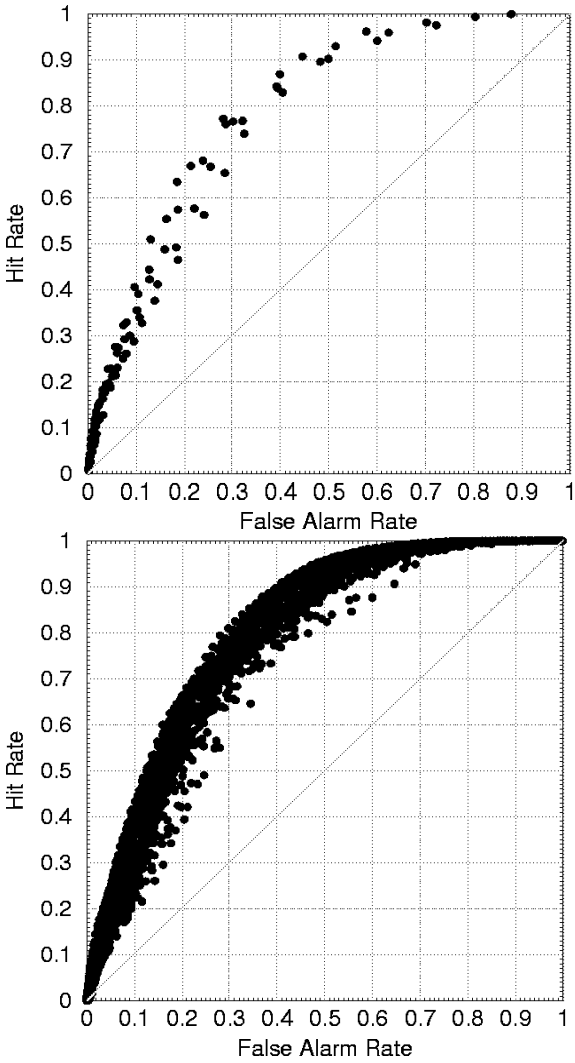


Fig. 7. Relative Operating Curve (ROC) for the 5 mm/12 h observed threshold, based on the spatial multi-event contingency tables shown in Table 5 and Table 6. **(a)** ECL (ECMWF T159 model, +72 h to +96 h); every point indicates the performance of a deterministic forecast based on the fact that the model forecasts at least 1, 2, 3, etc., mm/12 h at less than 100, 200, etc., km from the considered location (20 × 5=100 points). **(b)** ECEPS (ECMWF EPS, +72 h to +96 h); every point indicates the performance of a deterministic forecast based on the fact that at least 1, 2, 3, etc., ensemble members forecast at least 1, 2, 3, etc., mm/12 h at less than 100, 200, etc., km from the considered location (51 × 20 × 5=5100 points).

2.7 Significance tests

Figure 8 is an example of a comparison between two curves of value obtained from spatial multi-event contingency tables based on different forecasting systems. Some differences appear for a wide range of C/L ratios. Are these differences statistically significant? This question is particularly important when verifying the performance of forecasting systems with respect to extreme events, such as intense precipitation that rarely occurs in the data sample. Furthermore, the method of verification implies the use of a large number of forecast categories, which emphasizes the effect of insufficient sampling.

As pointed out by Hamill (1999), spatial correlation and the non-normality of errors make it difficult to use common

Table 6. Spatial multi-event contingency table based on ECEPS (ECMWF EPS, +72 h to +96 h) for the 5 mm/12 h observed threshold. $H_{j,k,l}(FA_{j,k,l})$: number of Hits (False Alarms) for more than j members forecasting more than k mm/12 h at less than $l \times 100$ km from the observation. $HR_{j,k,l} = H_{j,k,l}/\Sigma H_{j,k,l}$ (Hit Rate). $FAR_{j,k,l} = FA_{j,k,l}/\Sigma FA_{j,k,l}$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds) × 51 (ensemble members) × 5 (distances to the observation) = 5100

ECEPS 5 mm/12 h			Observed	Not observed
⋮	⋮	⋮	⋮	⋮
Forecast > k mm/12 h	At least by j forecasts	At less than 100 km from the observation	$H_{j,k,1}$	$FA_{j,k,1}$
Forecast > k mm/12 h	At least by j forecasts	⋮	⋮	⋮
Forecast > k mm/12 h	At least by j forecasts	At less than 500 km from the observation	$H_{j,k,5}$	$FA_{j,k,5}$
⋮	⋮	⋮	⋮	⋮

hypothesis tests (e.g. t test) for assessing the significance of weather forecasting verification results. Computer-based methods of hypothesis testing have been used in this study to evaluate the significance of the results. A resampling method has been systematically applied in order to estimate the probability that differences in the relative value between two forecasting systems could have been obtained by chance.

The method consists of the construction of an empirical distribution of differences that are not statistically significant (Hamill, 1999). The probability that the actual difference belongs to this distribution, i.e. that the difference is not significant, is then evaluated. This null distribution is obtained by comparing the relative value for every C/L ratio of two sets of independent forecasts that should perform identically. These two sets are generated 1000 times by randomly choosing the forecasts from either one or the other forecasting systems. Since it is very likely that the errors are spatially correlated, all local forecasts valid for a given 12 h period are considered together as a unique case. Temporal correlation of errors is also probable. In order to limit the dependencies, forecasts valid for the 12:00–00:00 UTC period (afternoon precipitations) and for the 00:00–12:00 UTC period (morning precipitations) have been considered separately, so that no consecutive 12 h periods can be found in the sample.

The different steps of the procedure are as follows: (i) contingency tables are computed for every 12 h period for system A and system B; (ii) the sample of 12 h periods is randomly halved into 2 sub-samples; (iii) the relative value is computed separately from each sub-sample using the contingency tables; (iv) the difference between the relative value of the 2 sub-samples is computed; (v) the procedure is iterated 1000 times from (ii) to (iv); (vi) the probability that the difference between the actual value of system A and the actual value of system B is significant is estimated from the empirical distribution obtained at the end of step (v).

3 Results

Intense precipitation events occur more frequently during the afternoon. The results presented in this section are based solely on 12:00–00:00 UTC precipitations. Unless otherwise stated, the lead-time is +84 h (precipitations accumulated from +72 h to +84 h).

3.1 Ensemble vs. single run

An important requirement for an ensemble is that it performs better than a control single forecast based on the same model. As mentioned in Sect. 1, the superiority of probabilistic forecasts based on EPSs over deterministic control forecasts has been demonstrated. In this subsection, the value of an ensemble is compared to that of the control forecast on the basis of spatial multi-event contingency tables. This means that the performance of a probabilistic forecast based on the ensemble is compared to that of a probabilistic forecast based on the control single run. The latter is designed to represent the probabilistic judgment of an operational forecaster using a single model forecast. The comparison of these forecasts is meant to indicate the usefulness of an EPS in an operational environment, with respect to QPFs.

3.1.1 ECMWF ensemble vs. control forecast

Figure 9 shows the relative value of the ECMWF EPS (ECEPS) and the control forecast (ECL) for the total range of C/L ratios (0 to 1). For each C/L, the statistical significance of the difference between the curves of value has been evaluated through the resampling procedure described in Sect. 2. For the 5 mm/12 h observed threshold (Fig. 9a), ECEPS and ECL perform similarly for C/L above the optimal value (that corresponds to the sample frequency of the event, i.e. 0.07 approx.). For lower thresholds, as small as 10^{-4}, the superiority of ECEPS over ECL is confirmed by the curves of value, but the 90% significance level is reached only for a proportion of C/L ratios. By contrast, ECEPS is

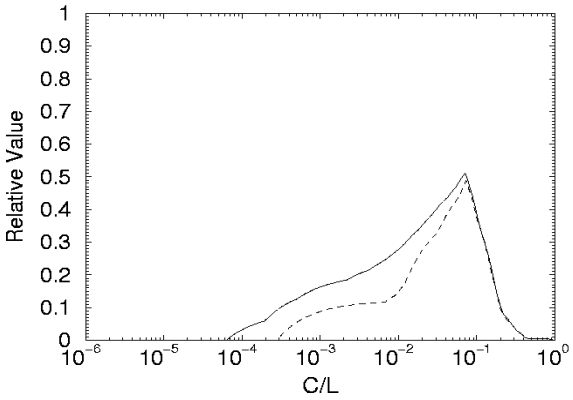


Fig. 8. Relative value for the 5 mm/12 h observed threshold, as a function of the user C/L, based on the spatial multi-event contingency tables shown in Table 5 and Table 6. Dash line: ECL (ECMWF T159 model, +72 h to +96 h). Solid line: ECEPS (ECMWF EPS, +72 h to +96 h).

significantly better than ECL with respect to the 20 mm/12 h threshold for a wide range of C/L ratios from approximately $2 \cdot 10^{-5}$ to 10^{-2} (Fig. 9b).

Although the 50 mm/12 h curves of value exhibit an advantage for ensemble forecasts for higher C/L ratios, significance tests show that ECEPS and ECL do not perform differently at the 90% level (Fig. 9c). Low significance of the results concerning the higher precipitation thresholds is probably due to, in this case as in many others presented below, the limited number of observed cases. For example, less than 100 cases of precipitations above 50 mm/12 h have been reported during the considered season. These 100 cases have occurred during 6 periods of 24 h, so that the number of independent observed cases is very small when only considering 12:00 UTC–00:00 UTC precipitations.

3.1.2 NCEP ensemble vs. control forecast

Figure 10 shows the relative value of the NCEP EPS running at 12:00 UTC (NCEPS12) and the corresponding control forecast (NC12). The differences are often not significant at the 90% level. When the differences are significant, NC12 is generally better than NCEPS12, except for smaller C/L ratios for 20 mm/12 h (Fig. 10b) and 50 mm/12 h (Fig. 10c). This result is rather disappointing, but indicates the importance of model resolution for QPF. The horizontal resolution of NC12 is T126, while the resolution of the 4 perturbed members of NCEPS12 is only T62.

The relative value of the NCEP EPS running at 00:00 UTC (NCEPS0) and the corresponding T62 control forecast (NC0) has been examined. The lead-time is +96 h, so that afternoon precipitations are considered (value is consequently lower than in the previous results where the lead-time is +84 h). Surprisingly, the result of the comparison (not shown) is similar to that obtained at 12:00 UTC, although NC0 and NCEPS0 have the same resolution (T62). This result seems

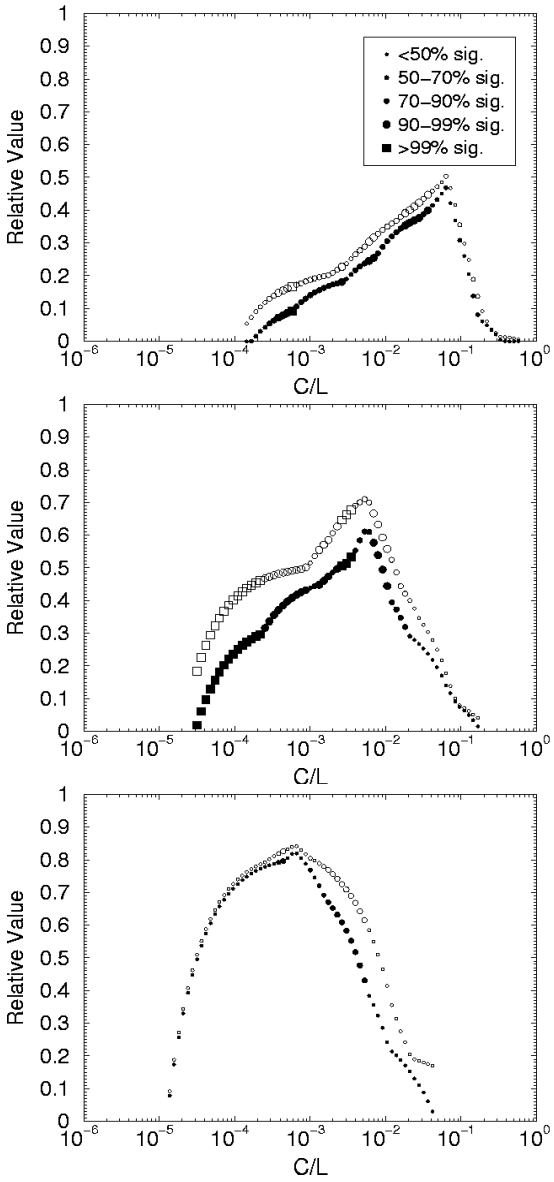


Fig. 9. Relative value as a function of the user C/L, based on spatial multi-event contingency tables. Afternoon precipitation only (+72 h to +84 h). Blank symbols: ECEPS (ECMWF EPS). Filled symbols: ECL (ECMWF T159 model). The size of the circles indicate the level of statistical significance of the difference in the value between the two forecasting systems: less than 50%, 50–70%, 70–90%, 90–99%. Squares indicate a level of significance above 99%. **(a)** 5 mm/12 h observed threshold. **(b)** 20 mm/12 h observed threshold. **(c)** 50 mm/12 h observed threshold. The legend indicated in Fig. 9a is valid for all figures from Fig. 9 to Fig. 16.

to contradict the conclusion of the previous paragraph about the importance of model resolution. It also contradicts the previous results based on 500 hPa geopotential height fore-

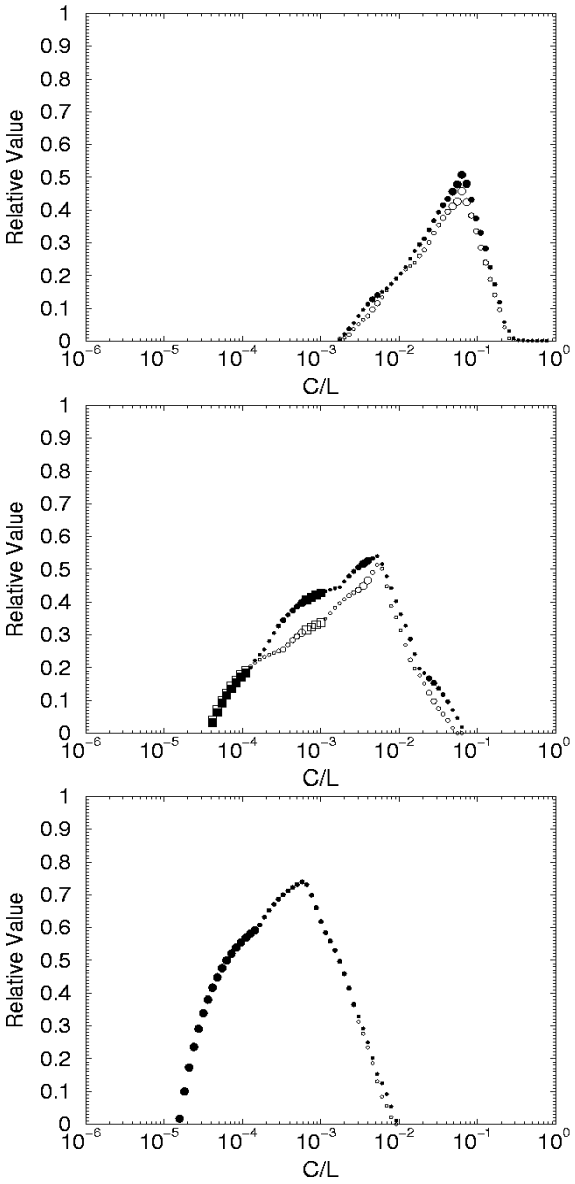


Fig. 10. Same as Fig. 9, but for NCEPS12 (NCEP EPS running at 12:00 UTC, blank symbols) and NC12 (NCEP T126 model, filled symbols). Differences are hardly visible, but there is a slight advantage for NCEPS12 in Fig. 10c.

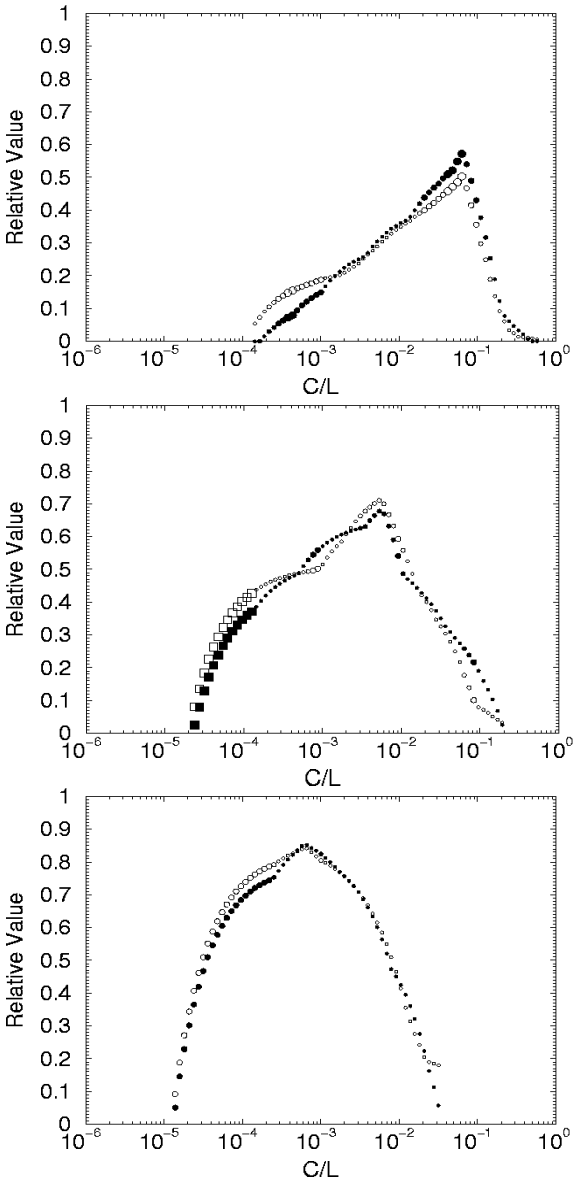


Fig. 11. Same as Fig. 9, but for ECEPS (ECMWF EPS, blank symbols) and ECH (ECMWF T319 model, filled symbols).

casts (Toth et al, 1998; Zhu et al, 2001). One should remain cautious in interpreting these contradicting results. The overperformance of the T62 control forecast might point to a special behaviour of the NCEP ensemble with respect to QPF. This possible weakness might be linked to essential differences in the ECMWF ensemble: (i) the method of generation of the perturbations; (ii) the lower resolution of the NCEP model; (iii) the limited population of the NCEP ensemble. The impact of (ii) and (iii) is examined in the next subsection.

3.1.3 ECMWF ensemble vs. higher resolution model single run

Figure 11 shows the relative value of the ECMWF EPS (ECEPS) and the higher resolution (T319) ECMWF model forecast (ECH). Most differences are not significant at the 90% level. For the 5 mm/12 h threshold (Fig. 11a), ECH is significantly better than ECEPS for the C/L corresponding to the maximum value, while the comparison is the opposite

for a lower C/L. For the 20 mm/12 h threshold (Fig. 11b), ECEPS is significantly better than ECH for C/L lower than 10^{-4} (at 99% level). No difference is significant for the 50 mm/12 h (Fig. 11c).

3.2 Ensemble vs. ensemble

In this subsection, ensembles running at 12:00 UTC have been compared in terms of the relative value computed from spatial multi-event contingency tables.

3.2.1 ECMWF ensemble vs. NCEP ensemble

Figure 12 shows the relative value of the ECMWF EPS (ECEPS) and the NCEP EPS running at 12:00 UTC (NCEPS12). ECEPS performs better than NCEPS12, with a high level of significance (often more than 99%) for the 5, 20 and 50 mm/12 h thresholds. This result is not surprising, given the difference in resolution of the models, on the one hand, and the difference in the number of members, on the other hand.

In order to evaluate the relative influence of these two factors, a smaller ensemble based on the ECMWF EPS has been constructed, consisting of the control forecast and the first 4 perturbed members. The value of this smaller ensemble (ECEPS5) has been compared to the value of the NCEP EPS (NCEPS12). The results (not shown) are very similar to those obtained with the fully populated ECMWF ensemble, indicating that the impact of the ensemble population might be much lower than the impact of the model resolution (or other characteristics of the ensembles, e.g. the method of generation of the perturbations).

3.2.2 ECMWF ensemble vs. smaller ensemble

In order to further investigate the impact of the ensemble population, smaller ensembles based on the ECMWF EPS control forecast and the 10/20/32 first perturbed members (ECEPS11, ECEPS21, ECEPS33) have been compared to the fully populated EPS (ECEPS). Note, however, that the first perturbed members' initial conditions are still obtained from all 25 singular vectors (SVs), since the perturbations are combinations of SVs (Molteni et al., 1996). This comparison thus favors smaller ensembles and only addresses the question of the number of integrations that are needed.

Value curves shown in Fig. 13 indicate that ECEPS21 performs as well as ECEPS, except for the 20 mm/12 h threshold and smaller C/L (order of 10^{-4}). Differences between ECEPS11 and ECEPS (not shown) are significant at the 90% level for a limited range of rather small C/L ratios for the 5 mm/12 h and 20 mm/12 h thresholds. No significant differences have been found for the 50 mm/12 h threshold and no significant differences have been found between ECEPS33 and ECEPS for any threshold (not shown).

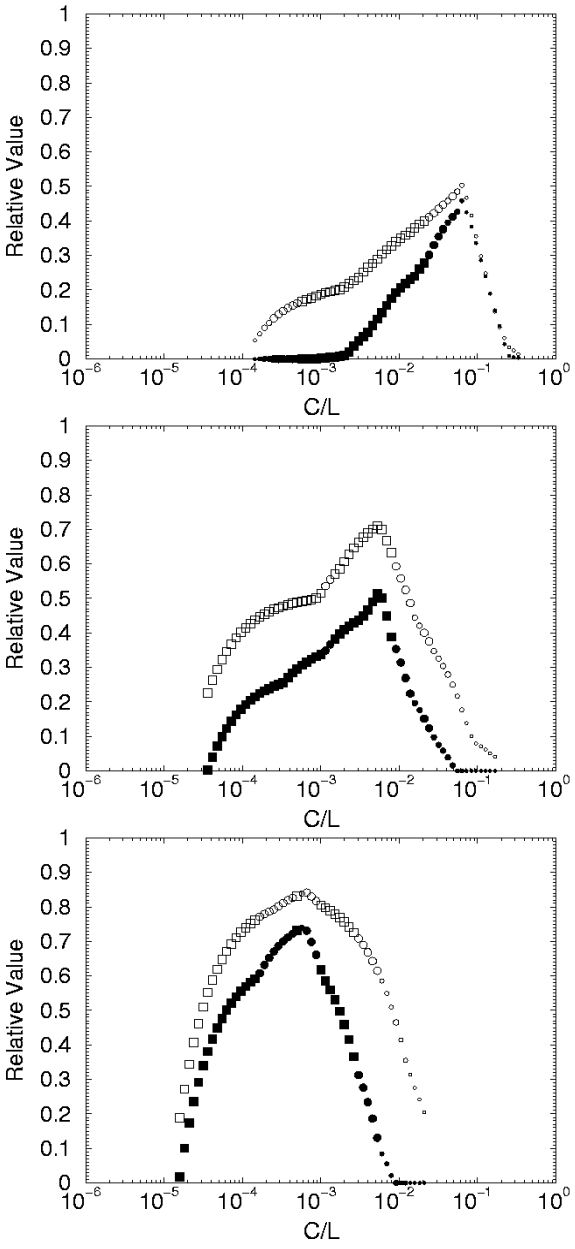


Fig. 12. Same as Fig. 9, but for ECEPS (ECMWF EPS, blank symbols) and NCEPS12 (NCEP EPS running at 12:00 UTC, filled symbols).

3.2.3 ECMWF ensemble vs. "poorman ensemble"

Since the impact of the ensemble population is limited, it might be interesting to consider the small "poorman ensemble" consisting of the ECMWF T159 control forecast and the NCEP T126 control forecast (ECNC). Since they simultaneously take into account the uncertainties of the initial conditions and model formulation, "poorman ensembles" have

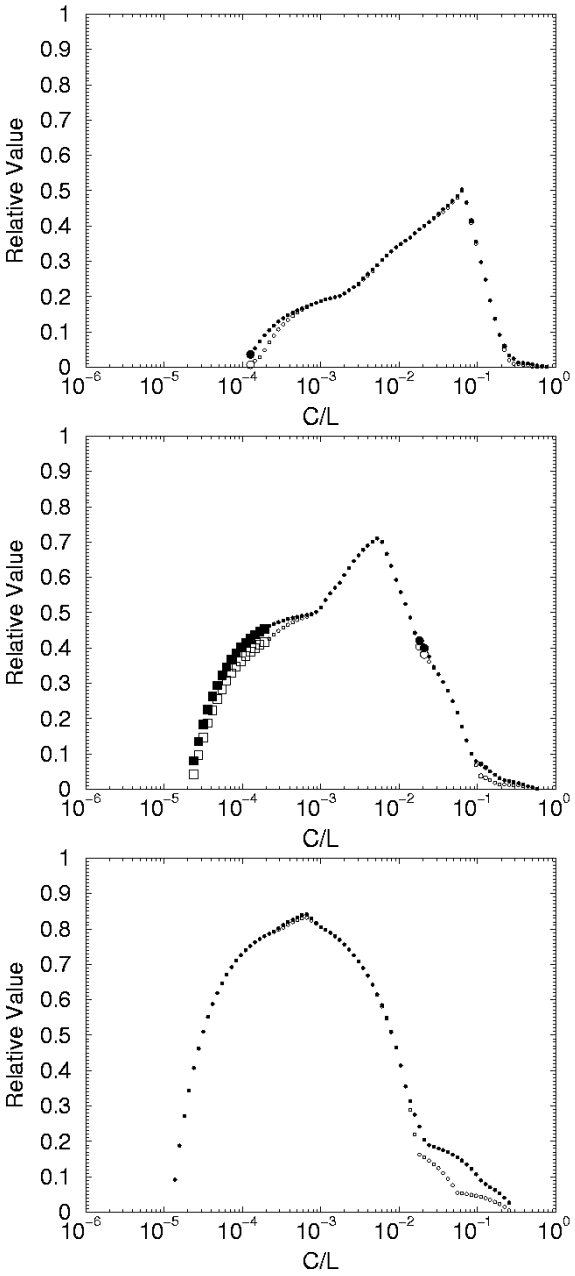


Fig. 13. Same as Fig. 9, but for ECEPS21 (ECMWF EPS control + 20 perturbed members, blank symbols) and ECEPS (ECMWF EPS, filled symbols).

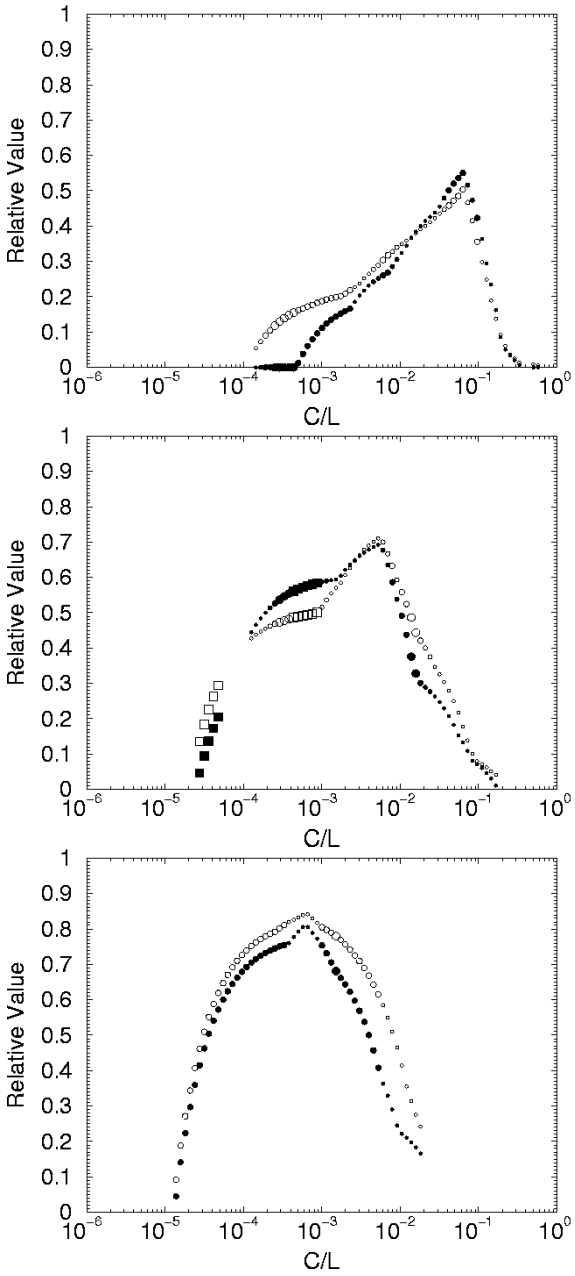


Fig. 14. Same as Fig. 9, but for ECEPS (ECMWF EPS, blank symbols) and ECNC ("poorman ensemble" consisting of the ECMWF T159 control forecast and the NCEP T126 control forecast, filled symbols).

proven to perform as operational EPSs for certain aspects of the prediction of upper level atmospheric parameters in the early medium-range (Ziehman, 2000; Atger, 1999).

Figure 14 shows the relative value of the ECMWF EPS (ECEPS) and the "poorman ensemble" (ECNC). Most differences are not significant at the 90% level. Significant differences indicate a superiority of ECEPS for smaller C/L ratios for 5 mm/12 h (Fig. 14a) and 20 mm/12 h (Fig. 14b). For the 20 mm/12 h threshold, ECNC is significantly better than ECEPS (at the 90% level) for a limited range of C/L ratios of the order of 10^{-3} (Fig. 14b). For the 50 mm/12 h threshold, ECEPS seems better than ECNC for any C/L, but no difference is statistically significant (Fig. 14c).

4 Discussion

4.1 Morning precipitations vs. afternoon precipitations

The results presented in the previous section only concern afternoon precipitations. Although the number of observations of intense precipitation is lower, the performance of ensembles and single model runs has been investigated with respect to the 00:00–12:00 UTC precipitations. Although most results are similar to those obtained for afternoon forecasts, some results differ with respect to comparisons between single models and EPSs.

Figure 15 shows the relative value of the ECMWF EPS (ECEPS) and the higher resolution (T319) ECMWF model forecast (ECH) for morning precipitations (+84 h to +96 h forecasts). In contrast with Fig. 11 (afternoon precipitations), ECEPS performs significantly better than ECH (at the 90% level) for a wide range of C/L ratios, especially for higher thresholds (20 mm/12 h and 50 mm/12 h). Similarly, the NCEP EPS running at 12:00 UTC (NCEPS12) performs as well as or even slightly better than NC12 (90% level significance for lower C/L ratios and for 50 mm/12 h), despite the higher resolution of the latter (T126) (not shown). As a consequence, ECEPS performs significantly better than the 2-member "poorman ensemble", consisting of the control forecasts of ECMWF and NCEP ensembles (ECNC), especially for lower C/L ratios and higher thresholds (not shown).

The difference in performance between morning and afternoon precipitations might come from the fact that operational ensembles are more likely to overperform a single run at longer lead-times (for a model running at 12:00 UTC: +96 h for morning precipitations, +84 h for afternoon precipitations). However, this hypothesis is not supported by the performance of the NCEP ensemble running at 00:00 UTC (NCEPS0), mentioned in Sect. 3.1.2, for +96 h forecasts of afternoon precipitations. For longer lead-times, the relative value of probabilistic forecasts of intense precipitation decreases and becomes close to zero for most C/L ratios, so that the performance of the ensembles over single runs cannot be demonstrated with confidence. This is the main reason for the choice of the 72–96 h range for the results presented in the previous section, although operational ensembles have been designed for use from day 3 to day 10 (ECMWF), and beyond (NCEP).

One intrinsic difference between the morning and afternoon precipitations is the frequency of convective activity. In France, during the winter season (as considered in this study), convective precipitations occur most frequently during the afternoon. Important precipitations occurring before 12:00 UTC are likely to originate from large-scale systems, while they are often a consequence of small-scale, convective activity when they occur after 12:00 UTC. The pdf of the morning precipitations is, therefore, primarily associated with large-scale dynamics uncertainty, while the intensity and location of the afternoon precipitations is more often largely unpredictable with operational global models. Operational ensembles have been designed for estimating variations in large-scale dynamics predictability. Probabilistic forecasts based on an EPS are thus likely to perform better for morning (large-scale) precipitations than afternoon (small-scale) precipitations. On the other hand, probabilistic forecasts based only on a single run take into account local uncertainties related to the location and intensity of the precipitation. Therefore, it is not surprising that they are more efficient with respect to the afternoon (small-scale) precipita-

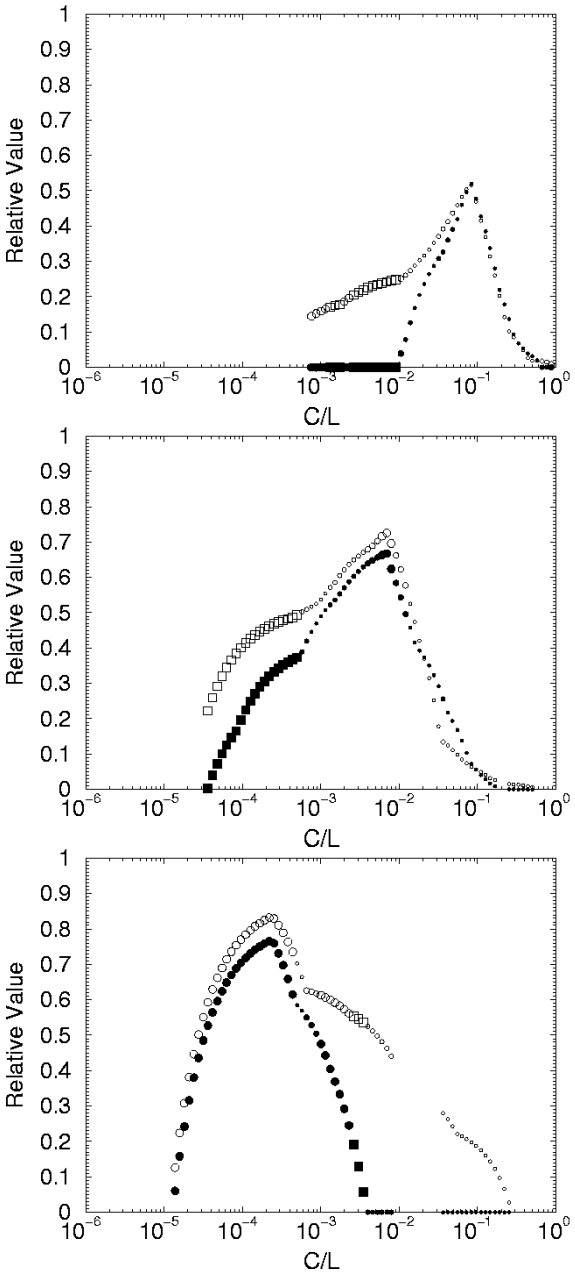


Fig. 15. Same as Fig. 11, but for morning precipitation (+84 h to +96 h).

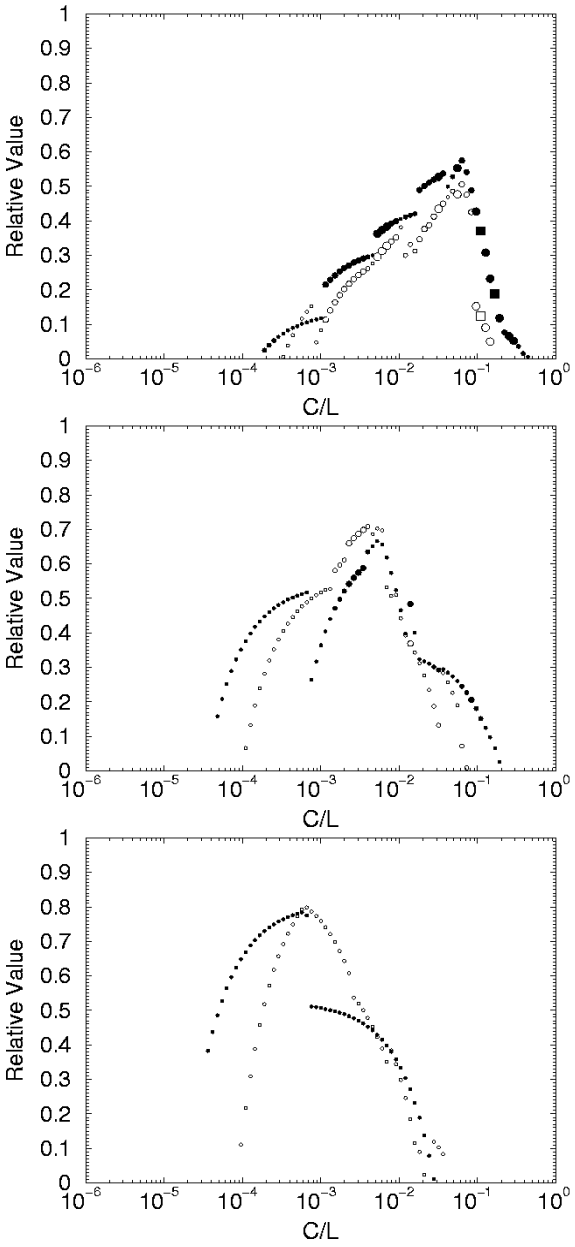


Fig. 16. Same as Fig. 11, but after halving the data in order to separate the sub-sample used for the derivation of the optimal forecast category for a given C/L, and the sub-sample used for verification.

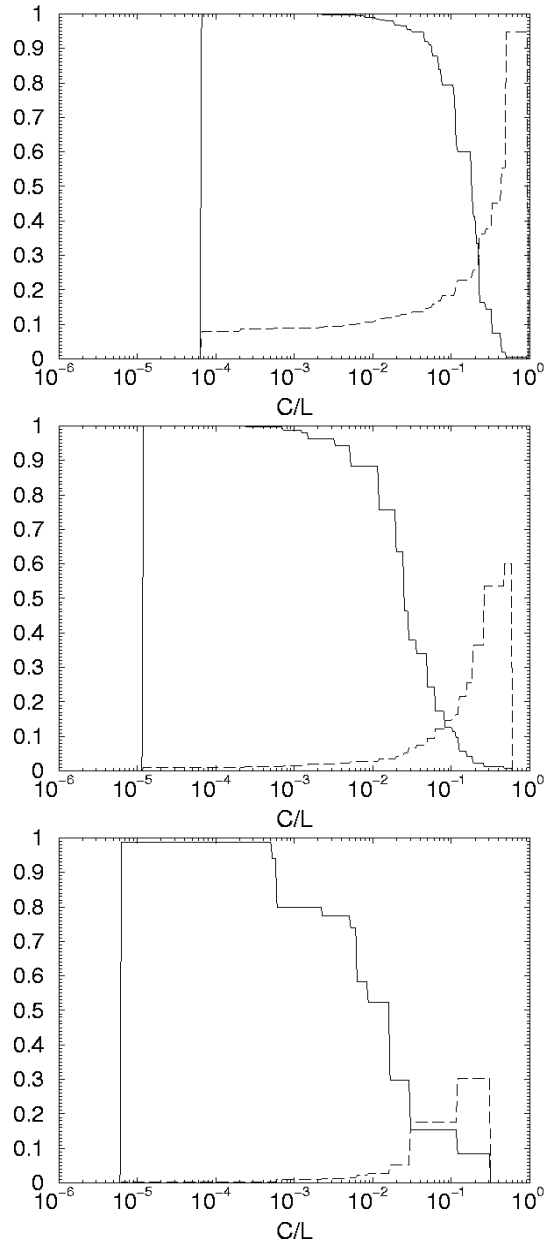


Fig. 17. Hit Rate (HR, solid line) and Correct Alarm Rate (CAR, proportion of forecasts of the event that are justified, dashed line) computed from spatial multi-event contingency tables based on the ECMWF EPS (ECEPS). For every C/L ratio, the HR and CAR are those obtained from the forecast category leading to the maximum value. Afternoon precipitation only (+72 h to +84 h). **(a)** 5 mm/12 h observed threshold. **(b)** 20 mm/12 h observed threshold. **(c)** 50 mm/12 h observed threshold.

tions, while uncertainties associated with large-scale systems are poorly estimated.

4.2 Computation of value from an independent sample

The results presented in the previous section have been obtained through an evaluation of probabilistic forecasts based on spatial multi-event contingency tables. As performed in most studies (e.g. Richardson 2000), the computation of the False Alarm Rate and Hit Rate, leading to the relative economic value of the forecast, has been performed under a

strong assumption: the forecast user is supposed to protect when the category forecast by the system leads to the maximum value that can be expected. For example, in the case of an EPS forecast evaluated from a standard contingency table, the user protects at least a certain number of ensemble members every time, with a forecast more than the considered threshold. The user can only know this certain number from the past. Proper evaluation should thus require an independent, representative sample, from which the optimal forecast category would be derived for every C/L. In practice, the available sample is generally small, so that it is used for both the evaluation and the determination of the optimal categories. One may qualify the result of this computation as a potential value (Richardson, 2000), i.e. the maximum value that is attainable in real conditions.

In order to evaluate the difference between the potential value and the actual value, the data have been randomly halved into 2 sub-samples. The first sub-sample is used for the derivation of the forecast category that leads to a maximum value for every C/L. The relative value is computed from the second sub-sample for the forecast category determined from the first sub-sample. Figure 16 shows the relative value of the ECMWF EPS (ECEPS) and the higher resolution (T319) ECMWF model forecast (ECH) when this procedure is followed. Most differences are not significant at the 90% level, except for the 5 mm/12 h threshold. The curves look rather noisy, with discontinuities reflecting the variability of the maximum value attained for a given C/L. This indicates that both sub-samples are too small (45 days each) to obtain conclusive results with respect to the actual value of the probabilistic forecasts of intense precipitations. When differences are significant, the actual value of the single forecast is higher. EPS forecasts probably suffer more than single forecasts, given the fact that the sample is small when compared to the number of forecast categories: $5 \times 20 \times 51$=5100 categories in the case of the EPS, but 5×20=100 categories for the single forecast. In other words, ensemble forecasts would have the potential to overperform single forecasts for the prediction of intense precipitations, but a larger sample would be needed in order to identify from past statistics the forecast category that leads, in effect, to the maximum value.

4.3 The meaning of very small C/L ratios

One of the aims of this study is to evaluate the usefulness of operational forecasting systems for the prediction of intense precipitations. The results presented in the previous sections show that the maximum potential value increases with the precipitation threshold. Impressively high levels of the potential value (80% of that attainable with a perfect forecast) are obtained for the 50 mm/12 h threshold. However, the range of users who benefit from intense precipitation forecasts is limited to very small C/L ratios. By construction, the maximum potential value is obtained for C/L ratios that are close to the frequency of occurrence of the event. The prediction of rare events thus is a benefit primarily to lower C/L users: those users facing a decision making situation that oblige them to protect (or take action, in a more general sense) as soon as the risk of potential damage exists, even if it is almost nil. This may be the case, for instance, of a mountaineer who requires a 99% probability of quiet weather, before deciding to go for a 3 day expedition in a remote area during winter.

When forecasting extreme events, the problem may just come from false alarms. The high maximum value obtained by operational forecasting systems for the prediction of rare precipitation events reflects the fact that high hit rates can be achieved provided that the frequency of false alarms is high. In practical terms, only well informed, professional users can tolerate a high frequency of false alarms. This is equivalent to saying that the C/L ratio of these users is small. By contrast, the majority of the users, especially among the public, hardly tolerate false alarms. The C/L ratio of these users is large, actually much larger than the climatological frequency of the considered event.

Figure 17 shows the hit rate and the Correct Alarm Rate (CAR, proportion of forecasts of the event that are justified) computed from spatial multi-event contingency tables based on the ECMWF EPS. For every C/L ratio, the HR and CAR correspond to the forecast category leading to the maximum value. Assuming that most users would require at least 30–50% of correct alarms, they could expect a 10–30% hit rate for 5 mm/12 h, but virtually no detection for 20 and 50 mm/12 h. This indicates that intense precipitation forecasts based on operational forecasting systems, although exhibiting high levels of maximum potential value, are only useful for a restricted category of users.

5 Summary

The performance of single models and ensemble prediction systems has been investigated with respect to quantitative precipitation forecasts, with a special emphasis on intense precipitation. Evaluation has been based on the relative economic value of the forecasts, computed from spatial multi-event contingency tables. A probabilistic forecast from an EPS can thus be compared to a probabilistic forecast based on a single model run. The latter is designed to represent the probabilistic judgment of an operational forecaster, from which any probabilistic or deterministic statement originates. The statistical significance of the comparisons between various forecasting systems has been estimated through a resampling procedure.

The relative value increases with the precipitation threshold. Impressively high levels of relative value (60–80% of that attainable with a perfect forecast) are reached for the 20 mm/12 h and 50 mm/12 h thresholds. These numbers reflect high hit rates that are obtained at the expense of a dramatic increase in the frequency of false alarms. The ECMWF EPS performs better overall than a single forecast based on the same model, even when the resolution of the ensemble is lower (T_L159 vs. T_L319). The difference is important for morning precipitation, especially for higher precipitation

thresholds and lower C/L ratios. On the other hand, the performance of the ECMWF EPS and single forecasts is rather similar for afternoon precipitation, probably due to more frequent convective events.

The NCEP EPS performs as well as a single forecast based on the same model for morning precipitation, even when the resolution of the ensemble is lower (T62 vs. T126). Higher resolution single forecasts perform better with respect to afternoon precipitation. The ECMWF EPS performs better than the NCEP EPS running at 12:00 UTC. This is still true when the population of the ECMWF ensemble is reduced to 5 members. More generally, the impact of reducing the number of members of the ECMWF EPS is rather small. No differences have been found between 51 and 33 members. The 11 member ensemble still performs as well as the fully populated ensemble for a limited range of C/L ratios. A "poorman ensemble", consisting of the control forecasts of the ECMWF and the NCEP EPSs, performs as well as the ECMWF EPS for afternoon precipitation. The ECMWF EPS is still significantly better with respect to morning precipitation.

Acknowledgement. C. Ziehmann, D. Richardson, B. Houdant and O. Talagrand, as well as an anonymous reviewer, provided helpful comments on an earlier version of this manuscript. Part of this work has been supported by ECMWF as part of an expert visit by the author in summer 2000.

References

Ångstrøm, A. K.: Probability and practical weather forecasting (in Swedish), Centraltryckeriet, Teknologföreningens Förlag, 11pp, 1919.

Atger, F.: The skill of ensemble prediction systems, Mon. Wea. Rev., 127, 9, 1941–1953, 1999.

Buizza, R. and Palmer, T. N.: Impact of ensemble size on ensemble prediction, Mon. Wea. Rev., 126, 2503–2518, 1998.

Buizza, R., Hollingsworth, A., Lalaurette, F., and Ghelli, A.: Probabilistic predictions of precipitations using the ECMWF Ensemble Prediction System, Wea. Forecasting, 14, 168–189, 1999.

Buizza, R., Petroliagis, T., Palmer, T. N., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., and Wedi, N.: Impact of model resolution and ensemble size on the performance of an ensemble prediction system, Quart. J. Roy. Meteor. Soc., 124, 1935–1960, 1997.

Carter, G. M., Dallavalle, J. P., and Glahn, H. R.: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system, Wea. Forecasting, 4, 401–412, 1989.

Chessa, P. A. and Lalaurette, F.: Verification of the ECMWF Ensemble Prediction System forecasts: a study on synoptic patterns, to Wea. Forecasting, 16, 611–619, 2001.

Courtier, P., Freyder, C., Geleyn, J. F., Rabier, F., and Rochas, M.: The Arpege project at Meteo-France, in: Proceddings on Numerical Methods in Atmospheric Models, 2, (Eds) ECMWF, 192–231, 1991.

Hamill, T. M.: Hypothesis tests for evaluating numerical precipitation forecasts, Wea. Forecasting, 14, 155–167, 1999.

Katz, R. W. and Murphy, A. H.: Economic value of weather and climate forecasts, (Eds) Cambridge University Press, 1997.

Liljas, E. and Murphy, A. H.: Anders Ångstrøm and his early papers on probability forecasting and the use/value of weather forecasts, Bull. Amer. Meteor. Soc., 75, 1227–1236, 1994.

Mason, I.: A model for assessment of weather forecasts, Aust. Met. Mag., 30, 291–303, 1982.

Mason, S. J. and Graham, N. E.: Conditional probabilities, Relative Operating Characteristics, and Relative Operating Levels, Wea. Forecasting, 14, 713–725, 1999.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Quart. J. Roy. Meteor. Soc., 122, 73–119, 1996.

Murphy, A. H.: The value of climatological, categorical and probabilistic forecasts in the Cost/Loss situation, Mon. Wea. Rev., 105, 803–816, 1977.

Murphy, A. H.: Probabilistic weather forecasting, in: Probability, Statistics, and decision making in the Atmospheric Sciences, (Eds) Murphy, A. H. and Katz, R. W., Westview Press, 337–377, 1985.

Murphy, A. H.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Wea. Forecasting, 8, 281–293, 1993.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Mon. Wea. Rev., 115, 1330–1338, 1987.

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., and Tribbia, J.: Ensemble prediction, in: Proceedings of Validation of Models over Europe (Eds) ECMWF, 1, 21–66, 1993.

Richardson, D. S.: Skill and economic value of the ECMWF Ensemble Prediction System, Quart. J. Roy. Meteor. Soc., 126, 649–668, 2000.

Simmons, A. J., Burridge, D. M., Jarraud, M., Girard, C., and Wergen, W.: The ECMWF medium-range prediction models development of the numerical formulations and the impact of increased resolution, Meteorol. Atmos. Phys., 40, 28–60, 1989.

Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of common verification methods in meteorology, WMO/WWW Tech. Rep. 8, 1989.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proceedings of Predictability, (Eds) ECMWF, 1–25, 1997.

Toth, Z. and Kalnay, E.: Ensemble forecasting at NCEP and the breeding method, Mon. Wea. Rev., 125, 3297–3319, 1997.

Toth, Z., Zhu, Y., Marchok, T., Tracton, S., and Kalnay, E.: Verification of the NCEP global ensemble forecasts, Preprints, 12th Conf. on Numerical Weather Prediction, Phoenix, Arizona, Amer. Meteor. Soc., 286–289, 1998.

Tracton, M. S. and Kalnay, E.: Operational ensemble prediction at the National Meteorological Center: practical aspects, Wea. Forecasting, 8, 379–398, 1993.

Ziehmann, C.: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models, Tellus, 52, 280–299, 2000.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: On the economic value of ensemble based weather forecasts, Bull. Amer. Meteorol. Soc., submitted, 2001.

Chapitre 3

Variabilité spatiale et temporelle de la fiabilité de prévisions probabilistes issues d'un ensemble Conséquences pour la calibration

Autour de l'article : "Spatial and interannual variability of the reliability of ensemble based probabilistic forecasts - Consequences for calibration" (Atger, 2003)

1. Introduction

Dans les procédures d'évaluation de la performance des prévisions numériques, on distingue traditionnellement le biais (erreur moyenne) et l'écart type d'erreur. Ces deux indicateurs de performance sont simplement les deux premiers moments de la distribution des erreurs. L'hypothèse sous-jacente est que la distribution des erreurs est normale, ce qui est une approximation très acceptable quand la quantité à laquelle on s'intéresse a le bon goût de présenter une distribution climatologique à peu près normale. Biais et écart type sont aussi les 2 termes d'une des nombreuses décompositions possibles de l'erreur quadratique moyenne (EQM), indicateur de qualité très populaire dans la communauté météorologique.

Le score de Brier (1950) est lui aussi une EQM (cf. chapitre 1 pour les définitions), mais la décomposition traditionnelle sous la forme biais^2+variance n'est pas utilisée, tout simplement parce qu'elle ne permet pas de mettre en évidence les caractéristiques essentielles de la qualité d'une prévision probabiliste (et aussi probablement parce que la quantité prévue n'est pas distribuée suivant une Gausienne). La décomposition courante du score de Brier permet de mettre en évidence un terme de fiabilité et un terme de résolution. Le terme de fiabilité n'est rien d'autre que la moyenne quadratique d'un biais conditionnel. Evaluer le degré de correspondance entre probabilité prévue et

fréquence observée (quand cette probabilité est prévue) revient en effet à comparer une probabilité prévue p à la "probabilité" moyenne observée q quand la probabilité p est prévue.

Le biais conditionnel est peu utilisé pour les variables météorologiques "classiques" (géopotentiel, température) pour lesquelles on se contente généralement du couple biais / écart type. Il l'est davantage pour des variables dont on sait que la distribution climatologique s'éloigne fortement de la normale. C'est le cas par exemple de la quantité de précipitations prévues, pour laquelle on a coutume de calculer un "biais de fréquence" pour un certain nombre de classes arbitrairement définies, par exemple 0-0.3 mm, 0.3-2 mm, 2-5 mm, etc. Pour chacune de ces classes, on compare les fréquences prévue et observée. On met ainsi en évidence la tendance du modèle à sous-estimer ou surestimer la fréquence d'occurrence de telle ou telle classe. On calcule parfois, même si cela est peu fréquent, la moyenne quadratique des biais de prévision dans chaque classe. Un tel indicateur correspond très précisément à ce qu'indique le terme de fiabilité du score de Brier pour une prévision probabiliste.

Parce qu'il s'agit d'un biais, le terme de fiabilité du score de Brier n'est informatif qu'à la condition d'être estimé à partir d'un échantillon homogène. Comme pour n'importe quel biais de prévision, les cas de surestimation de la probabilité prévue sont en effet susceptibles de compenser les cas de sous-estimation, pouvant ainsi conduire à une sous-évaluation du biais, c'est à dire à une surévaluation de la performance. Par exemple, si sur une région du globe la probabilité de précipitation est surestimée sur les océans et sous-estimée sur les continents (ce qui est souvent le cas en pratique), alors il serait dommage de calculer la fiabilité sur la totalité de la région sans tenir compte de la différence de performance du système de prévision entre la terre et la mer. C'est pourtant ce qui est fait couramment, non seulement dans la littérature (par exemple : Buizza et al., 1999) mais aussi pour l'établissement des scores qui sont rendus publics

par les centres opérationnels (voir par exemple sur le site du CEPMMT: http://www.ecmwf.int/products/forecasts/d/guide/verification/eps/proba_T850).

La fiabilité n'est pas seulement variable dans l'espace, mais aussi dans le temps. Il y a d'abord des variations saisonnières : une surestimation de la probabilité de précipitation peut apparaître sur les continents en saison chaude, du fait de la convection. Il y a aussi des variations d'une année à l'autre, en raison des variations climatiques, mais aussi parce que les ensembles changent, ainsi que les modèles sur lesquels ils sont construits. A nouveau, une évaluation peut ne pas être pertinente, c'est à dire peut ne pas conduire à des résultats à haute valeur informative, si des périodes de surestimation systématique de la probabilité prévue viennent compenser des périodes de sous-estimation systématique, entraînant une sur-évaluation de la fiabilité.

Une stratification des données est par conséquent nécessaire pour obtenir des évaluations pertinentes de la fiabilité. D'un autre côté, les données disponibles sont en quantité limitée, et cette stratification risque de conduire à travailler sur des échantillons petits. Le terme de fiabilité du score de Brier étant la moyenne quadratique des biais de chaque catégorie de probabilité, il est susceptible d'être surévalué lorsque certaines catégories sont sous-échantillonnées. On peut s'en convaincre en observant n'importe quel diagramme de fiabilité obtenu à partir d'un petit échantillon (par exemple : Talagrand et al., 1997) : la courbe de fiabilité est manifestement "bruitée" par le sous-échantillonnage, et l'écart d'une telle courbe à la diagonale (écart qui représente graphiquement le terme de fiabilité) est évidemment plus important que si la courbe était plus "lisse", c'est à dire si les catégories étaient suffisamment échantillonnées pour que la correspondance entre probabilité prévue et fréquence observée soit significative.

On met ainsi en évidence le paradoxe qui préside à l'estimation du terme de fiabilité du score de Brier : considérer de gros échantillons, susceptibles de ne pas être homogènes, conduit à une sous-évaluation du biais, tandis que la stratification risque d'entraîner une surévaluation en raison du sous-

échantillonnage. Les échantillons disponibles étant déjà de taille limitée (du simple fait du coût élevé de production) et les possibilités de stratification étant virtuellement illimitées, un compromis est à trouver.

Dans l'exemple mentionné ci-dessus d'une prévision de l'occurrence de précipitation, il apparaît clairement que la fiabilité dépend (au moins en partie) des biais systématiques qui affectent le modèle sous-jacent. Si la probabilité de précipitation est surestimée sur les océans, ou bien sur certaines régions continentales en saison convective, c'est évidemment parce que le modèle génère trop de phénomènes précipitants. Le biais du modèle n'est pas seul en cause, puisque la probabilité prévue dépend du caractère plus ou moins sous-dispersif de l'ensemble, mais il contribue à dégrader la fiabilité. Curieusement, le lien entre fiabilité de l'ensemble et biais du modèle, et plus généralement entre performance de l'ensemble et défauts systématiques du modèle, est resté largement ignoré dans les premiers temps de la validation des ensembles. C'est l'utilisation du diagramme de rang (ou diagramme de Talagrand) qui a permis que ce lien soit mis en évidence dans différents travaux. Tandis que le caractère sous-dispersif de l'ensemble se traduit par une forme en U typique (la vérification se trouvant souvent en dehors de l'intervalle prévu), le biais systématique du modèle a pour effet de rendre le diagramme de rang dissymétrique (la vérification se trouvant systématiquement décalée d'un côté de l'intervalle prévu). En pratique, les diagrammes sont souvent "en crochet", montrant l'effet conjoint du biais du modèle et de la sous-dispersion de l'ensemble, entre autres causes possibles (Hamill, 2001).

De la même façon qu'on corrige les biais systématiques des modèles de prévision en appliquant des méthodes statistiques (dites d'adaptation) on peut améliorer la fiabilité de prévisions probabilistes issues d'un ensemble par la calibration. La méthode la plus simple consiste à répercuter sur les probabilités prévues les surestimations ou sous-estimations constatées dans le passé. Par exemple, si dans le passé la fréquence observée d'un événement était de 30% lorsque 40% des

membres de l'ensemble prévoyaient son occurrence, on peut appliquer la règle consistant à prévoir une probabilité de 30% lorsque 40% des membres prévoient l'événement (Zhu et al., 1996). Lorsqu'on s'intéresse à la prévision d'une quantité scalaire, telle que la température ou la quantité de précipitations, des méthodes de calibration plus sophistiquées (mais pas forcément plus efficaces) existent pour corriger la distribution prévue par l'ensemble de manière à la rapprocher de la distribution conditionnelle des observations, améliorant ainsi globalement la fiabilité de la prévision de tous les événements conditionnés par la quantité considérée (Hamill & Colucci, 1998).

Concrètement, la calibration doit conduire à une amélioration de la fiabilité qui soit perceptible par l'utilisateur : celui-ci étant destinataire de prévisions locales, c'est la fiabilité locale qui doit être améliorée, et non la fiabilité telle qu'elle peut être estimée à partir de données provenant de vastes domaines sur lesquels la performance du modèle et de l'ensemble sont susceptibles de varier fortement. L'hypothèse sous-jacente à la calibration (comme à toute méthode d'adaptation statistique) est celle de l'identité des caractéristiques statistiques entre l'échantillon d'apprentissage, qui sert à définir la règle de correction, et l'échantillon d'évaluation sur lequel on applique cette règle. Une calibration s'appuyant sur une évaluation de la fiabilité qui ne tiendrait pas compte des variations locales serait efficace, mais peu pertinente puisque susceptible de n'avoir qu'un effet limité sur la fiabilité locale. On peut tenir un raisonnement similaire pour ce qui concerne la variabilité interannuelle de la fiabilité : il est inutile de constituer un échantillon d'apprentissage sur de longues périodes du passé si les caractéristiques du système (modèle et ensemble) ont changé, que ce soit à la suite de modifications apportées par ses concepteurs ou du fait de la simple variabilité climatique.

Du raisonnement qui précède on pourrait conclure qu'il suffit de constituer des échantillons parfaitement homogènes, c'est à dire qui ne mélangent ni les points de prévision, ni les saisons, ni les années. On pourrait même aller au-delà et ne

considérer ensemble que des cas qui présentent des similitudes météorologiques suffisamment marquées pour assurer une parfaite identité statistique entre échantillons d'apprentissage et d'évaluation. Ce serait un principe applicable si des données étaient disponibles en très grand nombre, mais ce n'est évidemment pas le cas. En pratique, les échantillons disponibles sont très limités. On l'a vu ci-dessus, la stratification des données est susceptible de conduire à un sous-échantillonnage des catégories de probabilité qui aurait pour effet de limiter l'efficacité d'une procédure de calibration. Ce risque est nettement moins important quand on considère une procédure de correction de biais, puisqu'il n'y a pas de catégorisation préalable à l'apprentissage. La stratification permet plus facilement de constituer des sous-échantillons d'apprentissage qui soient à la fois homogènes et de taille suffisante pour espérer obtenir une correction efficace du biais.

Une hypothèse assez naturelle est que la variabilité spatiale et temporelle de la fiabilité d'un ensemble est en grande partie une conséquence de la variabilité spatiale et temporelle du biais du modèle. Par conséquent il pourrait être judicieux de faire précéder la calibration d'une correction du biais afin d'utiliser au mieux les données disponibles en évitant le sous-échantillonnage. Une correction statistique du biais, à partir d'une stratification des données disponibles, pourrait conduire à une forte réduction de la variabilité de la fiabilité. L'apprentissage pour la calibration pourrait alors se faire sur la totalité des données disponibles. Cette approche a été proposée initialement à l'occasion d'un atelier du CEPMMT (Atger, 1999). Elle a été reprise et appliquée avec un certain succès au Met'Office, pour la prévision probabiliste opérationnelle (Mylne et al., 2001).

2. Méthode (section 2)

Après une présentation des données utilisées on rappelle brièvement la décomposition du score de Brier proposée par Murphy (1973), déjà introduite dans le chapitre 1 de la thèse. Suit une discussion sur la manière de catégoriser

les probabilités prévues pour effectuer cette décomposition (cette discussion est reprise et largement approfondie dans le chapitre 4 de la thèse). Il existe bien-sûr une catégorisation "naturelle", la seule à permettre une décomposition exacte du score de Brier, qui consiste à ne regrouper que des probabilités strictement égales, de $p=0/n$ à $p=n/n$ (si n est le nombre de membres de l'ensemble). Cependant, quand n est grand (ce qui est de plus en plus souvent le cas pour les ensembles opérationnels), cette catégorisation requiert de très gros échantillons pour que toutes les catégories soient représentées par un nombre suffisant de cas. Pour cette raison, les catégories sont regroupées de manière arbitraire dans la plupart des évaluations, le plus souvent sous la forme d'intervalles de 10% (Mullen & Buizza, 2001). La catégorie [p=0] est parfois distinguée car elle est souvent la plus peuplée, surtout quand on considère des événements peu fréquents. En pratique, le nombre de cas regroupés dans de telles catégories varie beaucoup, les faibles et fortes probabilités étant plus souvent représentées (ceci étant par ailleurs un signe de qualité de l'ensemble). Il y a donc un risque de voir certaines catégories insuffisamment peuplées pour obtenir un résultat significatif. La méthode proposée dans cette étude consiste à regrouper les cas [p=0] (catégorie insécable) puis à répartir les cas restants de manière à assurer un équilibre optimal entre 4 autres catégories.

La technique utilisée pour la calibration est ensuite décrite en détail. Il s'agit d'appliquer la méthode classique basée sur l'analyse de la correspondance entre probabilité prévue et fréquence observée. Cependant, les catégories de probabilité étant définies en fonction de l'échantillon, comme expliqué ci-dessus, il n'est pas possible de calibrer directement à partir de la fréquence observée pour une catégorie de l'échantillon d'apprentissage (sauf pour la catégorie [p=0] bien-sûr). La méthode proposée consiste en une interpolation linéaire entre les fréquences observées des différentes catégories.

La correction de biais consiste à translater la distribution de l'ensemble en fonction de l'erreur systématique de la moyenne d'ensemble. Cette correction fait

appel à une régression linéaire. Dans la littérature, le terme de régression n'est généralement pas utilisé pour corriger le biais, on se contente du terme additif. La raison en est qu'on souhaite éviter la tendance des valeurs corrigées à se rapprocher de la normale climatologique (la célèbre "regression to the mean" décrite par Pearson). Dans le cas où on corrige une distribution d'ensemble, cette tendance n'a pas beaucoup d'importance puisque la dispersion de l'ensemble permettra que des valeurs extrêmes soient prévues de toute façon.

3. Principaux résultats (section 3)

3.1. Variabilité inter-annuelle de la fiabilité

Dans cette sous-section on met en évidence les variations de la fiabilité en fonction de l'hiver considéré (Fig. 4). Bien qu'un test statistique non paramétrique (proche de celui utilisé dans le chapitre 2 de la thèse, voir en annexe) indique que les différences entre les termes de fiabilité du Brier score sont peu significatives, on montre que la variabilité inter-annuelle devient négligeable après calibration, à condition que l'apprentissage soit effectué sur le même hiver que l'évaluation. Le même résultat est obtenu en appliquant une correction du biais, ce qui semble indiquer que la majeure partie de la variabilité inter-annuelle de la fiabilité provient des variations du biais du modèle d'une année à l'autre.

3.2. Variabilité spatiale de la fiabilité

Dans cette sous-section les 4 hivers sont regroupés et on met en évidence les variations de fiabilité en fonction du point de grille considéré (Fig. 7). Un test statistique est à nouveau utilisé (voir en annexe), cette fois pour montrer que près de la moitié des points ont une fiabilité significativement différente de celle d'un "point moyen" virtuel. Une calibration locale, c'est à dire pour laquelle l'apprentissage est effectué séparément en chaque point de grille, conduit à une forte amélioration de la fiabilité (Fig. 11). En ce qui concerne la correction du biais, une correction locale n'apporte pas grand chose par rapport à une correction pour laquelle l'apprentissage est effectué sur l'ensemble des points. Ce

résultat étonnant est interprété comme la conséquence de la forte variabilité inter-annuelle du biais.

3.3. Variabilité inter-annuelle de la fiabilité locale

Les deux aspects de la variabilité sont considérés simultanément dans cette sous-section. L'effet d'une calibration locale est plutôt négatif lorsque l'apprentissage est effectué sur le même hiver que l'évaluation (Fig. 12). Ce résultat décevant est interprété comme une conséquence du trop petit nombre de cas disponibles pour l'évaluation. Comme discuté en introduction à ce chapitre, la forte stratification des données conduit à une efficacité réduite de la procédure de calibration. En revanche, une correction locale du biais dans les mêmes conditions (apprentissage effectué sur le même hiver que l'évaluation) conduit à une nette amélioration de la fiabilité (Fig. 15). Ce résultat confirme qu'une part significative de la variabilité spatiale de la fiabilité provient des variations du biais du modèle d'un point à un autre. Ce qui reste de variabilité spatiale de la fiabilité, indépendamment du biais, ne peut être corrigé par une calibration locale du fait du trop petit nombre de cas considérés. Une calibration dont l'apprentissage est effectué tous points confondus, appliquée après correction locale du biais, ne permet pas d'améliorer davantage la fiabilité. Ce dernier résultat pourrait indiquer, contre toute attente, que la fiabilité d'un ensemble présente des spécificités locales marquées, indépendamment des variations locales du biais du modèle. L'efficacité des procédures traditionnelles de calibration, dont l'apprentissage est effectué tous points confondus, tient principalement à la correction du biais. Après correction locale du biais, seule une calibration locale conduirait à une amélioration supplémentaire de la fiabilité, si la taille des échantillons d'apprentissage le permettait.

4. Discussion (section 4)

4.1. L'estimation de la fiabilité

On discute dans cette sous-section du compromis qu'il faut trouver entre un gros échantillon hétérogène, entraînant la sur-évaluation de la fiabilité, et un petit

échantillon homogène, entraînant une sous-évaluation de la fiabilité. Une solution, suggérée par un réviseur de l'article, serait que les centres opérationnels fournissent des jeux de données suffisants pour permettre l'évaluation a priori de la performance des systèmes de prévision avant leur mise en service. Une autre possibilité serait de regrouper des points ou des stations où le comportement statistique de l'ensemble est similaire, de façon à accroître la taille de l'échantillon d'apprentissage.

4.2. Significativité des résultats

On tente dans cette sous-section de montrer ce que pourraient être les résultats si les données étaient disponibles en plus grand nombre. L'expérience consiste à utiliser 2 catégories de probabilités prévues au lieu de 5. Une autre expérience a été menée, avec des résultats similaires, consistant à augmenter la fréquence d'occurrence de l'événement considéré (0.5 écart type au lieu de 1 écart type). La plupart des résultats sont similaires. La principale différence est qu'on obtient un effet positif de la calibration locale, confirmant le rôle joué par le sous-échantillonnage dans l'expérience de base. Le caractère intrinsèquement local de la fiabilité est également confirmé, puisque seule une calibration locale permet d'améliorer encore la fiabilité après correction locale du biais (cf. 3.3).

4.3. Variations du biais et de la fiabilité

On note que les variations du biais du modèle sont encore plus marquées quand on considère des paramètres de surface ("weather parameters") et qu'on peut donc s'attendre à ce que la fiabilité de l'ensemble soit encore plus dépendante des biais pour ce qui concerne la prévision de ces quantités. On remarque également que les changements apportés à l'EPS du CEPMMT entre 1996 et 2000, qui sont supposés avoir apporté une amélioration de la performance de l'ensemble, ont eu un impact sur la fiabilité qui est au mieux du même ordre de grandeur que la variabilité interannuelle.

4.4. Conséquences pour la calibration

On discute dans cette sous-section des implications que ces résultats peuvent avoir pour la calibration des prévisions probabilistes issues des ensembles. La principale limitation de l'étude est que la performance des procédures de calibration et de correction du biais est fortement idéalisée. En effet, la partition des données disponibles implique une relation statistique optimale entre échantillons d'apprentissage et échantillons d'évaluation. Une expérience a été conduite afin d'évaluer ce que pourrait être l'efficacité de ces procédures dans des conditions réelles. Il s'agit également de justifier a posteriori le choix de la partition idéalisée, solution retenue pour obtenir des résultats exploitables. L'expérience "réaliste" conduit en effet à des résultats très affectés par la petitesse des échantillons considérés. D'une façon générale aucune procédure locale ne permet d'améliorer la fiabilité. Une combinaison de calibration et de correction de biais pourrait être la meilleure option.

Annexe

Description de la méthode utilisée pour tester la validité statistique de la variabilité inter-annuelle et spatiale

On a décrit au chapitre 2 (annexe) une méthode non paramétrique pour tester la significativité des différences de score entre deux systèmes de prévision. De manière similaire, on s'interroge ici sur la significativité des différences de fiabilité entre deux hivers (variabilité inter-annuelle, section 3.1) ou entre les différents points du domaine considéré (variabilité spatiale, section 3.2).

Pour la variabilité inter-annuelle le principe est le même qu'au chapitre 2 : il s'agit d'effectuer un regroupement des deux séries de prévisions issues des hivers A et B et d'en extraire, par tirage aléatoire, deux séries de prévisions qu'on considère issues des hivers virtuels A' et B'. La différence de fiabilité entre les hivers A' et B' n'a aucune raison, autre que le hasard de l'échantillonnage, d'être différente de 0. En effectuant un grand nombre de fois cette opération de tirage aléatoire on construit une distribution empirique des différences non significatives. En comparant cette distribution à la différence de fiabilité entre les hivers A et B on déduit une estimation de la probabilité que cette différence soit significative.

L'hypothèse d'indépendance entre les données constituant l'échantillon est supposée remplie. En toute rigueur, l'existence probable de corrélations temporelles et spatiales imposerait d'appliquer une correction au résultat du test.

Les étapes de la procédure sont les mêmes que celles décrites au chapitre 2, à la différence près que le calcul du terme de fiabilité du score de Brier est plus simple à mettre en œuvre que celui de la valeur relative :

1. On calcule la différence absolue entre les termes de fiabilité du score de Brier calculés séparément pour les hivers A (N_A cas) et B (N_B cas).

2. En regroupant les hivers A et B on obtient un ensemble de N_A+N_B cas.

3. On extrait par tirage aléatoire N_A cas de l'ensemble obtenu à l'étape 2, à partir desquels on calcule le terme de fiabilité du score de Brier pour un hiver virtuel A'. Le terme de fiabilité du score de Brier pour un hiver virtuel B' est calculé à partir des N_B cas restants. On calcule la différence absolue entre les termes de fiabilité pour les hivers A' et B'.

4. L'étape 3 est itérée 1000 fois. On obtient une distribution empirique des différences absolues entre les termes de fiabilité calculés sur 2 hivers A' et B' pour lesquels la fiabilité est équivalente.

5. La probabilité que la différence absolue entre les valeurs relatives des systèmes A et B (calculée à l'étape 1) soit significative est estimée à partir de la distribution obtenue à l'étape 4.

Pour tester la significativité de la variabilité spatiale, on calcule le terme de fiabilité du score de Brier en un point qui varie chaque jour par tirage aléatoire parmi les 48 points du domaine considéré. Cette opération est répétée 1000 fois, ce qui permet de construire une distribution empirique du terme de fiabilité en un point « moyen » virtuel. Cette distribution permet d'évaluer la probabilité que la fiabilité en chacun des points soit indiscernable de la fiabilité en un point moyen. A nouveau, l'existence probable de corrélations temporelles imposerait en toute rigueur une correction du résultat du test.

Bibliographie

Atger, F., 1999: Probabilistic forecasting at Météo-France. Proceedings, 7th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF, 53-56.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble based probabilistic forecasts - Consequences for calibration. *Mon. Wea. Rev.*, accepté.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Buizza, R., A. Hollingsworth, F. Lalaurette and A. Ghelli, 1999. Probabilistic predictions of precipitations using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168-189.

Hamill, T. M, 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.*, **129**, 3, 550-560.

Hamill, T. M., and S. J. Colucci, 1998: Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Mullen, S. L. and R. Buizza, 2001. Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-661.

Murphy, A. H., 1973. A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

Mylne, K., C. Woolcock, J.C.W. Denholm-Price and R.J. Darvell, 2001: Operational calibrated probability forecasts from the ECMWF ensemble prediction system - Implementation and verification. 8th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 60-66).

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings, Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts, 1-25.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conference on Weather Analysis and Forecasting, Norfolk, Virginia, Amer. Meteor. Soc., J79-J82.

Spatial and interannual variability of the reliability of ensemble based probabilistic forecasts
Consequences for calibration

Frédéric Atger, Météo-France

Accepted for publication in Monthly Weather Review

December 2002

Abstract

Reliability is an essential attribute of the quality of probabilistic forecasts. It indicates the correspondence between a given probability, and the observed frequency of an event in the case this event is forecast with this probability. The variability of the reliability of ECMWF ensemble forecasts has been investigated. Probabilistic forecasts of 850-hPa temperature anomaly have been considered during four consecutive winter seasons. Reliability appears highly variable in space and time. A proper evaluation thus requires a local, seasonwise verification, in order to avoid an overestimation of the performance. On the other hand, stratification of the data is likely to lead to an underestimation of the performance due to ill-sampling, so that a compromise has to be found.

Variations of model bias contribute for a major part to the interannual variability of the reliability. After bias correction, reliability is virtually constant during the considered period of four years, despite two major changes that have been implemented in the ECMWF EPS. Local variations of model bias contribute highly to the spatial variability of the reliability, but this contribution is only revealed when considering separate winter seasons.

Due to sampling limitations, local calibration and/or local bias correction are unlikely to bring a large improvement of reliability in operational conditions. A good scheme might rather be a combination of domain bias correction and domain calibration. Because calibration and bias correction require large samples of data for computing statistics, and at the same time other, independent samples are needed for validation, the implementation of an operational scheme for improving local reliability might be an arduous challenge.

1. Introduction

The validation of Ensemble Prediction Systems (EPS) has become an important field of research during the last few years. Since December 1992, both the U.S. National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-range Weather Forecasts (ECMWF) have produced operational forecasts based on ensemble prediction (Tracton and Kalnay 1993; Palmer et al. 1993). Several other centers worldwide have since implemented operational or quasi operational ensembles (e.g. Houtekamer et al. 1996; Kobayashi et al. 1996). Different strategies can be followed for setting up an operational ensemble system, e.g. model characteristics, number of ensemble members, method for generating perturbations. One of the goals of ensemble validation is to provide an estimation of the performance of the different options (Atger 1999). Another issue is the "return on investments" of an EPS, that may be illustrated by quantitative comparisons between the cost of running an ensemble and the economic value of the forecasts for the community of users (Richardson 2001).

An EPS is designed to provide an ensemble of realizations of the probability density function (pdf) of the future meteorological state. EPS verification thus consists in evaluating to which extent ensemble members properly sample this pdf. A large variety of scores have been used to estimate the performance of an EPS in that respect. Among these scores, the most widely used is the Brier score (Brier 1950), designed for quantifying the performance of a probabilistic forecast of a dichotomous event. The Brier score is simply the mean square error of forecast probabilities, under the assumption that a perfect probabilistic forecast is a deterministic forecast that proves correct. The Ranked Probability Score (Epstein 1969) is a generalization of the Brier score to scalar variables. The Continuous Ranked Probability Score (CRPS) allows an evaluation of the performance of a continuous probabilistic forecast (Bouttier 1994).

Two aspects of the performance of the probabilistic forecast of a dichotomous event can be distinguished. The first aspect is the correspondence between a given probability, and the observed frequency of an event in the case this event is forecast with this probability. The second aspect is the variability of the observed frequency of an event, when the forecast probability of this event varies. These two attributes, called respectively reliability and resolution, appear in the decomposition of the Brier score proposed by Murphy (1973). Hersbach (2000) and Talagrand and Candille (2001) have proposed two different decompositions of the CRPS in order to separate the contributions of reliability and resolution. In a more general sense, reliability and resolution can be seen as two aspects of the performance of an ensemble. Reliability consists in predicting a distribution that proves close, a posteriori, to the distribution of observations when this distribution is predicted. Resolution is the variability of the distribution of observations, when the predicted distribution varies.

A necessary condition for a good reliability is a flat rank histogram. Each bar of the histogram indicates the proportion of cases when the verification is found between two consecutive ensemble members, once all members have been sorted (Talagrand et al. 1997). Outliers are found in the two extreme categories, corresponding to those cases when the verification is outside the range of ensemble forecasts. Rank histograms computed from operational ensembles are generally U-shaped, which is traditionally interpreted as the consequence of the lack of ensemble spread that is a common characteristic of operational EPS (Anderson 1996). When considering surface or low level atmospheric parameters, rank histograms are not only U-shaped, they often exhibit an asymmetry due to numerical models systematic biases. For example, the rank histogram is strongly L-shaped when considering the temperature at 850 hPa in an area where the model suffers from a warm bias (Palany et al. 1999). Similarly, symmetric U-shaped rank histograms may be interpreted as a consequence of symmetric, conditional model biases, among other plausible explanations (Hamill 2001). A part of the lack of reliability of probabilistic forecasts based on an EPS might thus come from model systematic biases, rather than point to a weakness of the method used for generating the ensemble.

As noticed by Murphy (1993), the lack of reliability can be seen as a probability bias. Therefore, several authors have proposed a calibration of probabilistic forecasts in order to improve their reliability. Calibration consists in a statistical correction of the forecast probability, based on previous verification. For example, in the case of the probabilistic forecast of a dichotomous event, if the observed frequency in the past was .3 when the forecast probability was .4, the calibrated probability will be .3 when the raw probability is .4 (Zhu et al. 1996). Alternatively, a statistical correction can be applied to the ensemble distribution, in order to get it closer to the conditional distribution of observations (Hamill and Colucci 1998). This correction generally consists in inflating the distribution, i.e. increasing ensemble spread in order to flatten a U-shaped rank histogram. On the other hand, the lack of reliability is often partly due to model biases, in particular when considering surface or low level parameters. Therefore, applying a bias correction to individual forecasts might be required before attempting to improve reliability by correcting (i.e. generally increasing) the ensemble spread.

Scores that indicate the performance of operational forecasting systems are traditionally computed over large geographical domains, e.g. Europe or the U.S.A., in order to accumulate large samples from relatively short periods of time (typically one season). This is generally the case for the evaluation of operational EPS (Buizza et al. 1999a, Mullen and Buizza 2001), although some studies are based on local verification (Ziehmann 2000). This is also generally the case when a method of calibration of probabilistic forecasts is applied to an ensemble in order to improve reliability (Eckel and Walters 1998). Yet reliability may be highly variable in space, especially when considering surface

or low level parameters. As it is the case for any bias, the lack of reliability is likely to compensate from one point to another, e.g. because the probability is overestimated in some places while it is underestimated in other places. Local variations of model biases, due to dependencies to the model topography, are likely to contribute to this variability. Therefore the estimation of the reliability from a computation over large geographical domains may give a poor indication about the performance of the system. For the same reason, the positive impact of calibration might be meaningless when the correction is based on spatially averaged statistics. The usefulness of such a calibration scheme is rather limited in practice, since reliability may not be improved locally.

Ensemble reliability may strongly vary in time, too. Most variations of the performance are likely due to the variability of meteorological conditions, from one year to another, according to the season, and even at the synoptic time-scale. The variability of the reliability is also a consequence of modifications applied both to the model and to the EPS, affecting the quality of the model as well as of the initial generation of ensemble members. Evaluations of operational ensembles generally take into account the interseasonal variability of the performance by computing seasonal scores. On the other hand, one might be tempted to compute operational scores from several years of data in order to increase the size of the verification sample, especially when evaluating the local performance, rather than computing scores from large geographical domains. Similarly, local calibration might require statistics based on several years of data. Because of the variability of the performance from one year to another, reliability estimates might be unrealistic, as well as the positive impact of calibration.

The aim of the work described in this article has been to study the spatial and interannual variability of the reliability of ECMWF ensemble forecasts. Consequences for the calibration of probabilistic forecasts are also discussed. The impact of model biases is considered too. The methodology is described in section 2. Results are presented in section 3, discussed in section 4, summarized in section 5.

2. Methodology

2.1. Data

Four consecutive winter seasons (December to February) are considered, from 10 December 1996 to 28 February 2000. During these four years the operational version of the ECMWF EPS has been improved several times, as well as the model on which it is based, but the horizontal resolution remained the same (T_L159), as well as the number of ensembles members (control + 50).

All forecasts considered in the study are +96 h probabilities of the 850 hPa temperature anomaly being above 1 standard deviation at 48 grid points over Europe (35N, 60N, 10W, 25E, 5°/5° grid). The

climate reference, for the definition of anomalies, has been computed from the ECMWF 15-year re-analysis (Gibson et al. 1997). Three-month averages (December, January, February) have been computed at every grid point. The verification reference is the ECMWF high resolution analysis that was operational during the considered period (T_L319), interpolated at every grid point. Given the high density of upper air observations over the considered area (radiosondes, aircraft observations, etc.), these local interpolations are assumed to reflect the true state of the atmosphere (i.e. analysis error is close to zero). The standard deviation of the temperature anomaly has been computed at every grid point from the sample of analyses. Figure 1 shows that the frequency of occurrence of the considered event, i.e. a temperature anomaly above 1 standard deviation, is comprised between 0.148 and 0.194 over the domain.

The four winter seasons are verified either separately or together. Each winter season is randomly divided into 3 equally populated sub-samples. Each sub-sample is verified separately. Calibration and bias correction, when required, are based on statistics computed from two sub-samples and implemented in the remaining sub-sample. This verification scheme has been preferred to a simple split of the data in two equal halves in order (i) to increase the number of cases considered for computing calibration or bias correction statistics, and (ii) to avoid dependencies to intraseasonal changes, by randomizing the data inside a winter season.

2.2. Decomposition of the Brier Score

The Brier Score (BS) is defined as the mean square error of the probability forecast:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \qquad (1)$$

where N is the number of forecasts, p_i is the forecast probability, o_i is the verifying observation (1 if the event occurs, 0 if it does not) (Brier 1950). The Brier Score is easily transformed into a decomposition of 3 terms:

$$BS = \sum_{k=1}^{T}\frac{n_k}{N}(p_k - o_k)^2 + \sum_{k=1}^{T}\frac{n_k}{N}(o_k - o)^2 + o(1-o) \qquad (2)$$

when the sample of N forecasts has been divided into T categories, each comprising n_k cases of forecast probability p_k. o_k is the observed frequency when p_k is forecast. o is the observed frequency in the whole sample (Murphy 1973).

The first term of the decomposition is the reliability, i.e. the square difference between the probability and the observed frequency of the event, averaged over the T categories. The reliability term indicates to which extent the forecast probabilities are accurate (it is negatively oriented, as the

Brier score, i.e. the smaller the better). The second term is the resolution, i.e. the square difference between the observed frequency when p_k is forecast and the observed frequency for the whole sample. The resolution term indicates to which extent the different forecast categories do indicate different frequencies (it is positively oriented, i.e. the larger the better). The last term of the decomposition is the uncertainty, i.e. the variance of the observations. It indicates the intrinsic difficulty in forecasting the event and does not depend on the forecasting system.

2.3. Estimation of the reliability

The computation of the reliability from Eq. 2 requires a categorization of the forecast probabilities. In the case of ensemble based probabilistic forecasts, the decomposition of the Brier score is exact only when the categorization consists in n+1 categories from p=0/n to p=n/n, where p is the forecast probability and n is the number of ensemble members (Atger 1999). When evaluating large ensembles (e.g. n=51 for the ECMWF EPS) a large amount of cases is needed in order to sample the n categories. For this reason, reliability is often evaluated from a smaller number of probability categories, e.g. 10% probability intervals, i.e. from 0-10% to 90-100% (Mullen and Buizza 2001). This categorization leads to an approximate decomposition of the Brier score.

In practice, the number of cases in each probability category varies significantly. Specifically, higher and/or lower probabilities are more often forecast than probabilities lying close to the climate frequency. This is in fact a desirable feature of ensembles which contributes to improve some aspects of the performance (those aspects related to the resolution component of the Brier score). Some probability categories are thus likely to suffer from a lack of representative cases.

A user oriented evaluation implies that probability categories reflect the variety of probabilistic forecasts that have been delivered to forecast users. This implies the use of n+1 categories, as mentionned above, unless ensemble based probabilities have been rounded e.g. to 10% units. On the other hand, significant results are more likely obtained when probability categories are carefully designed in order to evenly distribute the cases among them, at least approximately, in order to avoid under-sampling as much as possible. This is obviously a serious concern when investigating the variability of the performance in space and time, that implies a strong stratification of the data. Therefore, in the present study, probability categories have been prescribed according to the latter option. Since the prior probability of the event, i.e. the frequency of occurrence over the considered sample, is rather small (approximately .175) the zero probability is most frequently forecast and defines the first category. Four other categories are defined so that the number of forecast cases is approximately the same in each of them. Probability categories thus differ according to the considered sample.

The solid curve in Fig. 2 is the so-called reliability curve, that indicates the correspondence between the forecast probability (x-axis), averaged over all the cases of the category, and the observed frequency (y-axis). All grid points are considered together, leading to the *domain* reliability (by contrast to the *local* reliability, computed at every grid point). The four winter seasons are considered together. The reliability component of the Brier score is 2.5 10^{-3}, it can be seen graphically as the weighted, averaged, squared distance between the reliability curve and the 45° line. The distribution of forecast probabilities is also shown (Fig. 2, inset). It indicates a rather balanced distribution of the cases among the 5 categories: approx. 60% in the first category (zero probability), approx. 10% in categories 2 to 5.

2.4. Calibration

Several methods have been proposed for the calibration of ensemble forecasts (e.g. Hamill and Colucci 1998). The method used in the present study is based on reliability diagram statistics (e.g. Zhu et al. 1996). In its simplest form, this method consists in applying the following rule: when the probability category *i* is forecast in the evaluation sub-sample, the calibrated probability is the frequency with which the event is observed in the calibration sub-sample when the *same* category *i* is forecast. As mentioned in 2.3, the choice of a "balanced" categorization implies that probability categories differ according to the considered sub-sample. Therefore it is not possible to calibrate directly the probability as indicated above. Three cases can be distinguished:

(i) No member forecast the event in the evaluation sub-sample. The calibrated probability is the observed frequency when no member forecast the event in the calibration sub-sample.

(ii) The number of members forecasting the event in the evaluation sub-sample is comprised between k_i and k_{i+1} (i=1 to 4) if k_i is the number of members forecasting the event, averaged over the cases belonging to the *i*-th category of the calibration sub-sample. The calibrated probability is linearly interpolated between f_i and f_{i+1}, if f_i is the observed frequency when the *i*-th category is forecast in the calibration sub-sample.

(iii) The number of members forecasting the event in the evaluation sub-sample is above k_5. The calibrated probability is linearly interpolated between f_5 and 1.

The following is an example for illustrating the calibration process in cases (ii) and (iii). Assume the 4th (resp. 5th) category of the calibration sub-sample consists of the cases when the number of ensemble members forecasting the event is comprised between 22 and 30 members (resp. between 31 and 51 members). The average number of members forecasting the event is k_4=25 in the 4th category, k_5=38 in the 5th category. Case (ii): if 27 members forecast the event in the evaluation sub-sample, then the calibrated probability is linearly interpolated between the observed frequencies f_4

and f_5 in the 4th and the 5th categories (since $k_4<27<k_5$). Case (iii): if 40 members forecast the event in the evaluation sub-sample, then the calibrated probability is linearly interpolated between the observed frequency f_5 in the 5th category and 1 (since $40>k_5$).

Note that it would have been more simple, in case (ii) above, to take f_4 as the calibrated probability (since 22<27<30). But this method would lead to use the same calibrated probability for a large number of different cases, while the relationship between the forecast probability and the observed frequency is more accurately taken into account through interpolation.

Figure 2 shows the effect of calibration to the reliability curve and to the distribution of forecast probabilities. All grid points are considered together for computing the calibration statistics, leading to a *domain* calibration (by contrast to a *local* calibration when calibration statistics are computed separately for every grid point). The four winter seasons are considered together, for calibration statistics as well as for the evaluation. The domain reliability after calibration is almost perfect, while the distribution of forecasts is little affected. The reliability term of the Brier Score is reduced from $2.5\ 10^{-3}$ to $1.7\ 10^{-5}$.

2.5. Correction of model biases

The correction of systematic model biases, when applied, consists in translating the ensemble distribution according to the average error of the ensemble mean. This allows the ensemble spread not to be affected by model bias correction. The procedure is the following:

(i) A linear regression is applied to the *correction sub-sample*. More precisely, the parameters α and β are defined such as to achieve a least-squares fit of the quantity $\alpha X+\beta$, where X is the ensemble mean forecast, to the verification.

(ii) This linear correction is applied to the evaluation sub-sample. The difference between the ensemble mean forecast before and after bias correction is then applied to individual ensemble forecasts:

$$x_c = x+(\alpha-1)X+\beta \qquad (3)$$

where x (resp. x_C) is the individual ensemble forecast, before (resp. after) correction.

This correction scheme has proven more efficient than a mere subtraction, from every ensemble member, of the mean error averaged over all ensemble forecasts. Although the main biases are corrected, this scheme is still very simple and limited in effect compared to more sophisticated procedures, e.g. Model Output Statistics (MOS) methods that would take into account statistical dependencies involving various parameters.

The effect of model bias correction to the domain reliability is shown in Fig. 3. All grid points are considered together for computing the correction statistics, leading to a *domain* bias correction (by contrast to a *local* bias correction when statistics are computed separately for every grid point). The four winter seasons are considered together, for computing statistics as well as for the evaluation. Figure 3a shows the impact to the reliability curve shown in Fig. 2. The reliability component of the Brier score is reduced from 2.5 10^{-3} to 0.4 10^{-3}, a substantial reduction but still much less than that obtained through domain calibration (Fig. 2). Figure 3b shows the rank histogram, before and after bias correction. Again, all grid points and the four winter seasons are considered together. A model cold bias is much probably the reason of the high proportion of verifications that are warmer than all ensemble members. After bias correction the histogram is almost symmetric, but the U shape indicates that reliability is still far from being perfect.

3. Results

3.1. Interannual variability

In this section all grid points are considered together, but the four winter seasons are considered separately, in order to investigate the interannual variability of domain reliability. The corresponding reliability curves are shown in Fig. 4. Table 1 (a) shows the reliability component of the Brier score averaged over the 4 winter seasons, as well as the interannual standard deviation. The latter is larger than the former, indicating a strong variability from one year to another. However, sampling limitations contribute to this variability. A computer-based method of hypothesis testing has been used to evaluate the significance of the differences between the 4 winter seasons. Differences between 1996-1997 and the other seasons are significant at the .98 level, while other differences are not significant at the .9 level.

The method for testing the significance of the difference between 2 winter seasons is based on the construction of an empirical distribution of differences that are presumably not significant. These differences are generated by computing the reliability component of the Brier score from 2 samples obtained by randomly choosing the data from either one or the other winter season. The probability that the actual difference between the 2 winter seasons belongs to this *nul* distribution is then evaluated (Hamill 1999). The different steps of the procedure are as follows. (i) Winter *A* and winter *B* cases are grouped together into a test sample. (ii) This test sample is randomly halved into 2 subsamples. (iii) The reliability component of the Brier score is computed separately from each subsample. (iv) The difference between the reliability component of the Brier score for the 2 sub--samples is computed. (v) The procedure is repeated 1000 times from (ii) to (iv). (vi) The probability that the difference between the actual reliability component of the Brier score for winter *A* and the

actual reliability component of the Brier score for system *B* is significant is estimated from the empirical distribution obtained at the end of step (v).

Consequences for calibration are shown in Fig. 5-6 and Table 1 (b) and (c). The effect of domain calibration is far from perfect when considering the four winter seasons together for computing the calibration statistics (Fig. 5). These results contrast with those shown in section 2.4, i.e. when the evaluation is done over the four winter seasons together. Fig. 5 looks like a mere translation of Fig. 4 and the interannual variability is unchanged (Table 1 (b)). On the contrary, computing calibration statistics from every winter season separately has a strong positive impact on domain reliability, since it reduces the average reliability as well as the relative variance (Fig. 6 and Table 1 (c)). Interannual variability is negligible after calibration. This can be seen as an a posteriori indication that variability before calibration is significant, despite the hypothesis testing results presented above.

The positive impact of domain bias correction is not as large as the improvement obtained through domain calibration, with respect to the average reliability component of the Brier score (Table 1 (d) and (e)). On the other hand, bias correction leads to a stronger reduction of the interannual variance, especially when computing correction statistics from every winter season separately (Table 1 (e)). Once bias correction has been applied, reliability is virtually the same for the four winter seasons. This indicates that a large part of the interannual variability of domain reliability might be a consequence of bias variations from one year to another.

3.2. Spatial variability

In this section the four winter seasons are considered together, but the grid points are considered separately, in order to investigate the spatial variability of the local reliability. Figure 7 shows the 48 local reliability curves corresponding to every grid point. The curves spread widely, indicating a large spatial variability. Figure 8 shows the distribution of the reliability component of the Brier score over the considered domain. Average local reliability is $3.9\ 10^{-3}$, i.e. 50% higher (i.e. worse) than domain reliability (see section 2.3). This indicates that domain evaluation may lead to overestimate the local performance with respect to reliability.

Although the categories have been defined carefully, as discussed in section 2.3, some curves in Fig. 7 appear rather noisy due to the relatively limited number of cases that have been considered at every grid point. With a sample of 351 days, each point of the curves is computed from only 35 forecast cases on average (10%), excepted the first point where the forecast probability is zero (60% of the cases). Therefore, the spatial variability shown in Fig. 8 may partly result from ill-sampling rather than indicating true discrepancies regarding the behavior of the forecasting system. For the same reason, the higher level of local reliability compared to domain reliability might be an artefact

due to sampling limitations. Figure 9 shows an estimate of the probability that reliability computed at a given grid point is significantly different from reliability computed at other grid points over the considered area. The method is similar to that described in section 3.1. The reliability component of the Brier score is first computed for a theoretical "average" grid point, varying randomly everyday from point 0 to point 47. This computation is repeated 1000 times, in order to construct an empirical distribution of the reliability of an "average" grid point. The actual reliability at every grid point is then compared to this distribution. Significance level is higher than .9 at 23 grid points out of 48, higher than .95 at 10 grid points corresponding to the minima and maxima of Fig. 8.

Consequences for calibration are shown in Fig. 10 and Fig. 11. Figure 10 (a) shows the spatial distribution of the local reliability after domain calibration, i.e. when all grid points are considered together for computing calibration statistics. The impact is almost uniformly positive, slightly detrimental at a small number of grid points (Fig. 10 (b)). The spatial structure of the field is almost unchanged and most local maxima are still present. Average local reliability is $2.0\ 10^{-3}$, i.e. reduced by a factor 2. Spatial standard deviation is reduced in the same proportion (from $2.9\ 10^{-3}$ to $1.5\ 10^{-3}$). By contrast, local calibration allows a stronger reduction of reliability, reflecting more closely the spatial distribution of reliability before calibration (Fig. 11). After local calibration no local maximum exceeds $2.5\ 10^{-3}$ and the spatial distribution is rather uniform (not shown). Average local reliability is $0.8\ 10^{-3}$, i.e. reduced by a factor 5. Spatial standard deviation is reduced in the same proportion (from $2.9\ 10^{-3}$ to $0.7\ 10^{-3}$).

The positive impact of domain bias correction is as large as the impact of domain calibration with respect to the average local reliability ($2.0\ 10^{-3}$). The reduction of the spatial variability is even larger (standard deviation is $1.3\ 10^{-3}$). Surprisingly, local bias correction brings only little further improvement to the average local reliability ($1.7\ 10^{-3}$). The spatial variability is reduced in the same proportion as it is after domain bias correction. This may indicate that local variations of model bias contribute little to the spatial variability of the reliability, when considering the four winter seasons together. This counterintuitive result might be a consequence of the large interannual variability of model bias. Further investigation has confirmed that the spatial variance of model bias is high, but varies a lot from one year to another (not shown).

3.3. Interannual variability of the local reliability

In the previous sub-section the strong improvement of local reliability after local calibration was obtained by considering the four winter seasons together. On the other hand, it was shown in section 3.1 that reliability is highly variable from one year to another. Therefore, a proper estimate of the local reliability, before and after calibration, would require a winterwise verification. The reliability component of the Brier score has been computed separately for the four winter seasons, at every

grid point. Table 2 indicates the interannual mean and standard deviation, after spatial averaging over all grid points. Before calibration, the spatial distribution for an average winter season is very much similar to the spatial distribution when the four winter seasons are considered together (not shown), but the average reliability is higher ($9.8\ 10^{-3}$ vs. $3.9\ 10^{-3}$).

The effect of a local, winterwise calibration is disappointing (Table 2 (c)). The impact is seriously detrimental on average at several locations (Fig. 12). Examination of the results for every winter season confirms that calibration is far from optimal (not shown). Surprisingly, the effect of calibration is slightly better on average when the four winter seasons are considered together for computing calibration statistics, the evaluation still being done winterwise (Table 2 (b)). The impact is positive almost everywhere (Fig. 13).

This last result seems to indicate that the interannual variability of the reliability would not affect the efficiency of a local calibration method. On the other hand, it might well be a mere consequence of ill-sampling. Local, winterwise calibration statistics, based on a small number of cases (2/3 of the season, i.e. 54 or 60 days depending on the winter season), might be not robust enough to represent a systematic behavior of the forecasting system. In order to investigate further this issue, reliability has been examined at grid point 16, located off Ireland, where the impact of local, winterwise calibration is detrimental on average. Figure 14 shows the reliability diagram at grid point 16, before and after calibration, for winter 1999-2000. Calibration based on the four winter seasons considered together has a limited positive impact overall, while the noisy reliability curve obtained after winterwise calibration leads to a large deterioration.

The impact of winterwise local bias correction appears much better than that of winterwise local calibration (Fig. 15). The reduction of the interannual variance of the reliability is also much larger (Table 2 (e)). Sampling limitations are apparently more detrimental for computing calibration statistics than bias correction statistics. This is probably because in the former case the sample has to be further stratified into a number of probability categories, while no further stratification is required in the latter case. The importance of interannual variations of model bias is confirmed by the relatively poor efficiency of a local bias correction considering the four winter seasons together for computing correction statistics (Table 2 (d)). Winterwise local bias correction leads also to better results than a local calibration considering the four winter seasons together for computing statistics, especially with respect to the reduction of the spatial variability of the performance (Table 2 (b)). The comparatively good efficiency of bias correction can be illustrated with the example of grid point 16 (Fig. 14): the last probability category is probably not sampled enough, but the rest of the curve indicates a rather good reliability.

One issue discussed at the end of section 3.2 is that local variations of model bias may contribute little to the spatial variability of reliability. Indeed a winterwise, domain correction of model bias has a substantial positive impact, similar to that obtained through winterwise, local calibration (Table 2 (f)). On the other hand, local correction has a larger impact on the average reliability (almost double than that obtained through domain correction) and allows a much stronger reduction of the interannual variance (Table 2 (e)). The spatial variance is reduced from 7.2 10^{-3} to 4.1 10^{-3} after local bias correction, for the average winter season. The impact to the spatial variance is almost as high after domain correction (4.9 10^{-3}). This seems to indicate that local statistics are not robust enough to allow a further reduction of the spatial variance, although reliability is improved at the local scale.

Further investigation has shown that applying local calibration after local bias correction leads to little further improvement, in terms of average reliability as well as the reduction of spatial and interannual variance. This is likely a consequence of the calibration scheme being much affected by sampling limitations. Applying domain calibration after local bias correction has not much effect either. This last result might indicate that ensemble reliability is intrinsically variable in space, independently of model bias.

4. Discussion

4.1. Estimation of the reliability

Because the reliability component of the Brier score is basically a squared bias, the noise introduced by under-sampling is likely to lead to underestimate the performance. On the other hand, it is common knowledge that biases tend to compensate one another when computed from an heterogenous sample, e.g. when considering different grid points or winter seasons together. Therefore, a compromise has to be found between (i) an overestimation of the performance caused by considering heterogenous data, and (ii) an underestimation of the performance caused by considering too small samples. This compromise is a general condition for evaluating properly the performance of a forecasting system, but it has a special importance in the case of the estimation of the reliability of ensembles forecasts, due to the necessary partitioning of the sample into probability categories.

From the results presented in section 3, one can conclude that the estimation of reliability requires a deep stratification of the data, in order to take into account spatial variability as well as interannual variability. Considering grid points and/or consecutive winter seasons together is likely to lead to estimates of the reliability that do not reflect the actual performance of the forecasts that are made available to the users (i.e. local forecasts whose reliability varies from one year to the next acording to interannual climate variations). On the other hand, a local, winterwise evaluation implies

inevitably a strong reduction of the available sample that may be detrimental for a proper evaluation. Even if one considers together consecutive winter seasons, data for local verification cannot be accumulated for long periods of time because ensembles and models change frequently, due to progresses in the field of numerical weather prediction. A period of four years, as considered in the present study, is probably a maximum (note that the model resolution of the ECMWF EPS was changed again in November 2000).

An ideal, definitive solution would be that operational centers provide long samples of forecasts performed with the current models and/or ensembles over past periods. Potential users could evaluate a priori the expected performance of operational forecasts. Calibration as well as any kind of statistical post-processing would be achieved in much better conditions than it is the case when working with small samples accumulated over short periods of time. However, the limitation of computer resources makes the implementation of such a practice very unlikely for the time being.

A possible strategy, that has not been explored further in the present study, might consist in estimating first the variations of local reliability, from a large sample considering several winter seasons together (or even different seasons). Locations (or stations, in case of weather parameters) might then be classified according to their similarities with respect to local reliability. A seasonwise evaluation might be conducted in a second step, considering several locations grouped together in order to increase the size of the sample.

4.2. Significance of the results

As discussed in section 2.3, the estimation of the reliability requires a careful definition of probability categories in order to avoid the effects of ill-sampling. In this study each category groups only 10% of the sample on average, except the first category. This leads to a strong further reduction of the computation sample, already stratified according to the winter season and the location: for a season of 90 days, 9 cases on average are considered in each category. On the other hand, these numbers depend obviously of both the number of probability categories (here, 5) and the parameter threshold that conditions the proportion of probability zero (here, 1 standard deviation).

In order to increase the number of cases considered in each probability category, only two probability categories have been used for verification in this section. This evaluation scheme can be seen as indicating the performance of a binary forecast: either the event is excluded (zero probability), or the event is possible (with a moderate probability, of the order of .3). The aim is to evaluate the potential impact of calibration and bias correction when sampling limitations are relaxed. Table 3 indicates the average reliability component of the Brier score, computed locally and winterwise, before and after calibration and/or bias correction. Most results obtained in the previous sections are

confirmed, in particular the impact of local model bias correction. Furthermore, the reduction of the spatial variance is much higher after local correction than after domain correction (not shown). This confirms that the limited impact of local bias correction is probably due to sampling limitations.

The main difference, compared to previous results, is the strong improvement obtained through local winterwise calibration. This is a confirmation that the poor efficiency obtained with 5 probability categories is the consequence of ill-sampling. On the other hand, the reduction of spatial variance is hardly larger than that obtained after local bias correction, confirming that much of the spatial variability of reliability is due to local bias variations. The reduction of interannual variance is also larger after local bias correction than after any kind of calibration. The largest impact overall is obtained when a local calibration is applied after local bias correction (both winterwise). On the contrary, applying domain calibration after local bias correction leads to no further improvement. This confirms that the spatial variability of ensemble reliability is not entirely due to model bias variations, as mentioned in section 3.3. There exist local discrepancies of the EPS statistical behavior, independent of the model bias, that result in local variations of the reliability.

Another modification has been tested, consisting in an increase of the frequency of occurrence of the considered event, so that the proportion of cases in each probability category is larger (due to the reduction of the frequency of the zero probability). The threshold has been reduced to 0.5 standard deviation, leading to an average frequency of occurrence around .3. Results (not shown) confirm those obtained when considering two probability categories instead of five for the evaluation.

4.3. Variations of model bias and reliability

The results presented in this study indicate that model bias contributes highly to the lack of reliability. It was shown that most of the improvement expected through calibration can be obtained after an adequate correction of model bias. Furthermore, the main part of the interannual variability of the reliability can be explained by the interannual variability of model bias. The spatial variability of model bias has also a large impact on local variations of variability. It should be kept in mind that the present study was conducted with a free atmosphere parameter (temperature at 850 hPa). Local variations of model bias are generally even larger when considering weather parameters (e.g. 2--meter temperature), and the impact of these variations to reliability is probably higher.

It was shown that reliability varies according to the considered winter season. This is mainly due to model bias variations, while the part of the reliability that is not related to model bias remains relatively constant during the considered period of four years. Two major changes have been implemented in the ECMWF EPS during the 1996-2000 period: introduction of evolved singular vectors (Barkmeijer at al. 1999) and simulation of random model errors (Buizza et al. 1999b). The

amplitude of the reliability improvement related to these changes, if any, appears much lower than the amplitude of interannual variations, most of these variations being due to systematic model biases.

4.4. Consequences for the calibration of probabilistic forecasts

Calibration and/or bias correction schemes have been used in this study in order to investigate the variability of reliability. From the results presented in previous sections, one might define a calibration strategy consisting in an optimal combination of bias correction and calibration. On the other hand, the results have been obtained with a partition of the data that has the effect of giving an ideal, unrealistic picture of the performance of any statistical correction. In the real world, winter statistics are not available before the end of the season, and the first month of the season cannot be calibrated and/or corrected from statistics computed from the next two months. Randomization of the data further increases the estimated performance of calibration and bias correction, since correlations between consecutive days emphasize the statistical consistency between the calibration (or correction) sub-sample and the evaluation sub-sample.

The conditions for a proper estimation of the reliability have been discussed in section 4.1. For defining a calibration strategy, one wants to estimate reliability as accurately as possible from available verification data. On the other hand, validation of the strategy would require an independent sample of data in order to test the efficiency of the method. In the real world, the size of available samples is dramatically limited. In this section the impact of calibration and bias correction has been tested from a different partition of the data, with the aim of giving a more realistic picture of the expected performance in an operational environment. In this new experiment, winterwise calibration or bias correction consists in computing statistics from the first half of the season (45 days) then applying them to the second half. Calibration or bias correction based on the four winter seasons considered together is replaced by computing statistics from one winter season, then applying them to the next winter season (this implying that the first winter season is not used for evaluation, while the last one is not used for computing statistics). Results (not shown) indicate a large effect of ill-sampling, for local as well as winterwise calibration. The effect of calibration is very limited overall, although domain statistics based on the previous winter season appears the most robust. Local bias correction also suffers much from ill-sampling, while winterwise domain statistics (i.e. based on the current winter season) lead to a moderate bias correction. The best combination might be a domain calibration based on the previous winter season, applied after domain bias correction based on the current winter season.

These results do not mean that local calibration and/or local bias correction are definitely useless. They still indicate that the positive effect of taking into account the spatial variability of the

reliability cannot be demonstrated from the available data. One might consider that the results obtained from an "idealized" partition of the data give sufficient evidence of the positive effect of local calibration and/or local bias correction, so that a calibration and/or correction scheme can be based on the whole available sample, with no need for further validation. Proper evaluation would come later, when new data are available. On the other hand, the fact that no evaluation is possible for a long time is hardly acceptable in an operational environment where the performance of the methods has to be monitored in real time.

Results related to the efficiency of calibration have been obtained by applying a simple, arbitrarily designed calibration method, described in section 2.4. Other methods exist that might give better results, in particular methods that consider the whole ensemble distribution instead of focusing on one peculiar probability threshold (Hamill and Colucci 1998). Correction of model biases has also followed a rather simple scheme, described in section 2.5. Statistical techniques traditionally used for post-processing numerical forecasts, e.g. Model Output Statistics (MOS) methods, might be able to reduce more efficiently model biases, including conditional biases. In particular, the combined effect of statistical adaptation and calibration, along the lines discussed in this article, is likely to improve significantly the reliability of probabilistic forecasts of weather parameters.

5. Summary and concluding remarks

Reliability is an attribute of the quality of probabilistic forecasts. It indicates the correspondence between a given probability, and the observed frequency of an event in the case this event is forecast with this probability. The variability of the reliability of ECMWF ensemble forecasts has been investigated. Probabilistic forecasts of 850-hPa temperature anomaly have been considered from December 1996 to February 2000. The four consecutive winter seasons have been verified both together and separately. Local reliability has been estimated at 48 grid points over Europe. Domain reliability, considering all grid points together, has also been estimated. The effect of model bias variations to the spatial and interannual variability of the reliability has been studied. The consequences of this variability for calibrating probabilistic forecasts have been examined too. The main conclusions of the study are the following:

1) Reliability is highly variable in space and time. A proper evaluation thus requires a local, seasonwise verification, in order to avoid an overestimation of the performance. On the other hand, stratification of the data implies a strong reduction of available samples, leading to an underestimation of the performance due to ill-sampling. A compromise has thus to be found in order to avoid these two pitfalls.

2) Variation of model bias from one year to another is the main source of the interannual variability of the reliability. After bias correction, reliability is virtually constant during the considered period of four years (only winter seasons are considered), although two major changes have been implemented in the ECMWF EPS in 1998.

3) Local variations of model bias contribute highly to the spatial variability of the reliability. However, the impact of these variations is only revealed when considering each winter season separately, due to the large interannual variability of model bias.

4) Seasonwise, local bias correction leads to a substantial improvement of reliability. Applying a calibration scheme after this bias correction brings only little further improvement. However, there exist local discrepancies of ensemble statistical behavior that contribute to the spatial variability of reliability independently of model bias local variations.

5) Due to sampling limitations, local calibration and/or local bias correction are unlikely to bring any significant improvement of the reliability in operational conditions. A good practical scheme might be a combination of domain bias correction (based on the current season) and domain calibration (based on the previous year).

6) Calibration and bias correction require large samples of data for computing statistics. At the same time, validation of the methods implies that other, independent samples are available. This double requirement means that the implementation of an operational scheme, along the lines discussed in this article, is not an easy task.

Most results presented in this work have been obtained from an evaluation of the efficiency of calibration and bias correction schemes applied to the idealized partition of the data described in section 2.1. As discussed in section 4.4 this evaluation gives an unrealistic picture of the performance of statistical corrections. The matter of the study is the variability of ensemble reliability, rather than the performance of statistical correction schemes. However, some credible results have been obtained about the *relative* efficiency of calibration and bias correction, according to the fact that grid points and/or consecutive winter seasons are grouped together or considered separately.

Sampling limitations that lead to pernicious effects would be relaxed if the considered event was more frequent and/or if the number of probability categories was smaller (as discussed in section 4.2). The choice of an anomaly above 1 standard deviation (prior probability around .17) is motivated by the fact that, in practice, forecasters express generally little interest in probabilities of events that occur more frequently (except forecasters in charge of seasonal forecasts). Forecasting rare events has always been more challenging than indicating a vague deviation from the normal.

Most standard ensemble products, as well as most evaluations of the performance of ensembles, focus on relatively unfrequent events.

The scope of this article might appear rather limited: a single lead time, a single quantity, a limited region. The choice of the temperature at 850 hPa is a compromise between the interest to end users and the availability of an acceptable reference for verification (i.e. a reliable analysis). This quantity is also subject to model bias characteristics that are close to those affecting surface "weather" quantities, e.g. 2-meter temperature. The choice of the lead-time and the region was arbitrary. Further work will be necessary to determine whether the results presented here can be generalized to other lead-times, other meteorological parameters, and other regions of the globe.

Acknowledgement. O. Talagrand and an anonymous reviewer provided helpful comments on earlier versions of the manuscript. Part of the work described in this article has been supported by ECMWF as part of an expert visit by the author in Summer 2000.

References

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate*, **9**, 1518-1529.

Atger, F., 1999. The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 9, 1941-1953.

Barkmeijer, J., R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2333-2351.

Bouttier, F., 1994: Sur la prévision de la qualité des prévisions météorologiques. Ph. D thesis. Université Paul Sabatier, Toulouse, France.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, N. Wedi, 1997: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **124**, 1935-1960.

Buizza, R., A. Hollingsworth, F. Lalaurette and A. Ghelli, 1999a. Probabilistic predictions of precipitations using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168-189.

Buizza, R., M. Miller, and T. N. Palmer, 1999b. Stochastic simulation of model uncertainties. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887-2908.

Eckel, F.A. and M.K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF Ensemble. *Wea. Forecasting*, **13**, 1132-1147.

Epstein, E.S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985-987.

Gibson, J. K., P. Kallberg, S. Uppala, A. Hernandez, A. Nomura, E. Serrano, 1997: ERA description. ECMWF Re-Analysis Project Report Series, vol. 1, 72 pp.

Hamill, T. M., and S. J. Colucci, 1998: Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Hamill, T. M, 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.*, **129**, 3, 550-560.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for Ensemble Prediction Systems. *Wea. Forecasting*, **15**, 5, 559-570.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie and H.L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Kobayashi, C., K. Yoshimatsu, S. Maeda, and K. Takano, 1996: Dynamical one-month forecasting at JMA. Preprints, 11th Conf. on Numerical Weather Prediction, Norfolk, VA, Amer. Meteor. Soc., 13-14.

Mullen, S. L. and R. Buizza, 2001. Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-661.

Murphy, A. H., 1973. A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

Palany. P, G. Richard, R. Verret, L. Lefaivre, G. Pellerin and P. Houtekamer, 1999: Ten day Temperature anomaly forecast based on ensemble prediction system, Proceedings of the 17th A.M.S. Conference on Weather and Forecasting, Denver, Co, 13-17 Sept. 1999.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tribbia, 1993: Ensemble prediction. Seminar on Validation of Models over Europe, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, vol. 1, 21-66).

Richardson, D. S, 2000. Skill and economic value of the ECMWF Ensemble Prediction System, *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.

Richardson, D. S., 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473-2489.

Talagrand, O., G. Candille, 2001: Evaluation of the continuous ranked probability score for probabilistic prediction of scalar variables. Abstracts of the XXVI European Geophysical Society General Assembly, Nice, France, March 2001.

Tracton M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting*, **8**, 379-398.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conference on Weather Analysis and Forecasting, Norfolk, Virginia, Amer. Meteor. Soc., J79-J82.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus*, **52**, 280-299.

Table 1. Reliability component of the Brier score, computed separately for the four winter seasons, mean and standard deviation. All grid points are considered together (domain reliability). Each winter season is weighted by the number of considered days (resp. 81, 90, 90 and 90).

	average	standard deviation
(a) No calibration - No bias correction	$29.1\ 10^{-4}$	$30.5\ 10^{-4}$
(b) Domain calibration (statistics based on 4 winters altogether)	$9.5\ 10^{-4}$	$10.6\ 10^{-4}$
(c) Winterwise domain calibration (statistics based on 4 winters separately)	$0.7\ 10^{-4}$	$0.4\ 10^{-4}$
(d) Domain bias correction (statistics based on 4 winters altogether)	$12.4\ 10^{-4}$	$8.7\ 10^{-4}$
(e) Winterwise domain bias correction (statistics based on 4 winters separately)	$2.9\ 10^{-4}$	$1.1\ 10^{-4}$

Table 2. Same as Table 1, but all grid points are considered separately then the local reliability is averaged over the 48 grid points. The effect of latitude is not taken into account.

	average	standard deviation
(a) No calibration - No bias correction	$9.8\ 10^{-3}$	$4.0\ 10^{-3}$
(b) Local calibration (statistics based on 4 winters altogether)	$6.3\ 10^{-3}$	$2.0\ 10^{-3}$
(c) Winterwise local calibration (statistics based on 4 winters separately)	$7.1\ 10^{-3}$	$1.0\ 10^{-3}$
(d) Local bias correction (statistics based on 4 winters altogether)	$7.6\ 10^{-3}$	$1.4\ 10^{-3}$
(e) Winterwise local bias correction (statistics based on 4 winters separately)	$5.6\ 10^{-3}$	$0.2\ 10^{-3}$
(f) Winterwise domain bias correction (statistics based on 4 winters separately)	$6.8\ 10^{-3}$	$1.0\ 10^{-3}$

Table 3. Same as Table 2, but only two probability categories have been used for the evaluation.

	average	standard deviation
(a) No calibration - No bias correction	$4.8\ 10^{-3}$	$3.8\ 10^{-3}$
(b) Local calibration (statistics based on 4 winters altogether)	$2.9\ 10^{-3}$	$2.0\ 10^{-3}$
(c) Winterwise local calibration (statistics based on 4 winters separately)	$0.8\ 10^{-3}$	$0.4\ 10^{-3}$
(d) Local bias correction (statistics based on 4 winters altogether)	$2.6\ 10^{-3}$	$1.2\ 10^{-3}$
(e) Winterwise local bias correction (statistics based on 4 winters separately)	$1.2\ 10^{-3}$	$0.0\ 10^{-3}$
(f) Winterwise domain bias correction (statistics based on 4 winters separately)	$1.9\ 10^{-3}$	$0.3\ 10^{-3}$

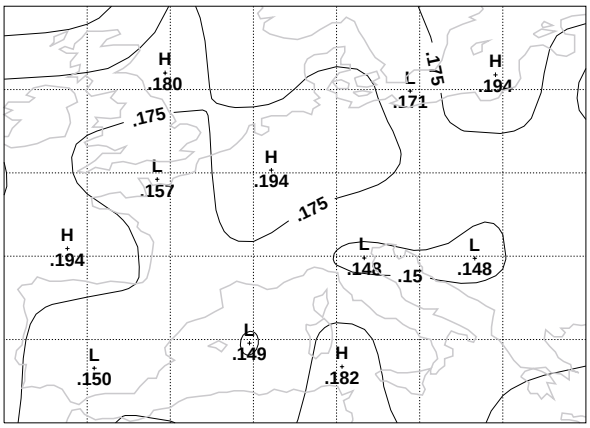


Figure 1. Frequency of occurrence of the considered event, i.e. 850 hPa temperature anomaly above 1 standard deviation, computed at every grid point from the whole sample of ECMWF analyses. Interval is .025.

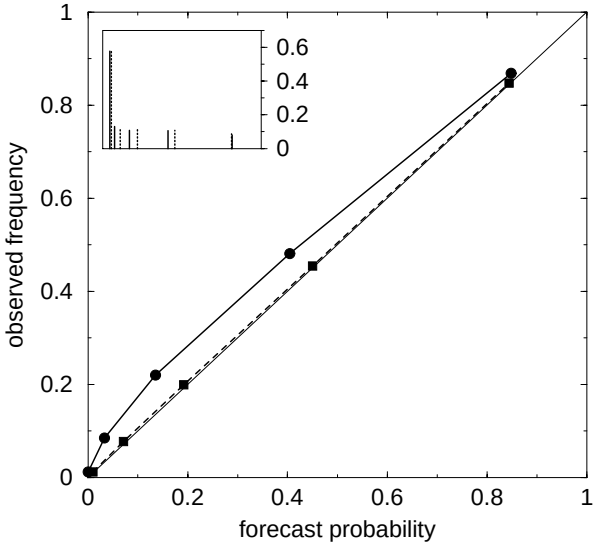
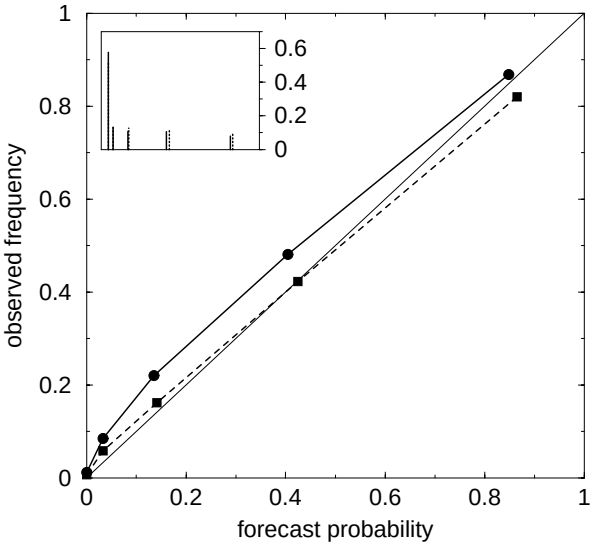
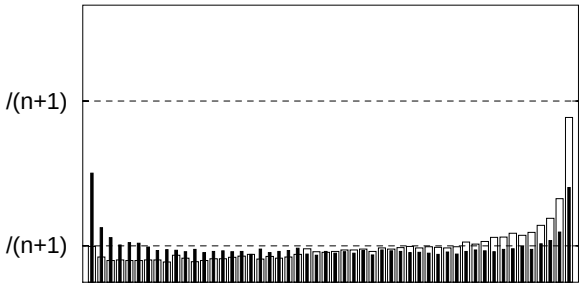


Figure 2. (main graph) Reliability diagram, before (solid line, circles) and after calibration (dashed line, squares). The frequency of observation when the forecast belongs to the category is plotted against the average forecast probability in the category. (inset) The frequency of forecasts in the category is plotted against the average forecast probability in the category, before (solid line) and after calibration (dashed line). All grid points and all winter seasons are considered together.



(a)



(b)

Figure 3. (a) Same as Fig. 2, but the dashed line indicates the effect of bias correction. (b) Rank histogram, before calibration (blank bars) and after calibration (black bars). All grid points and all winter seasons are considered together. Each bar of the histogram indicates the proportion of cases when the verification was found between two consecutive ensemble members, once all members have been sorted. Outliers are found in the two extreme categories. A perfect reliability would imply a flat histogram, with a proportion of cases 1/(n+1) in all categories where n is the number of ensemble forecasts (here n=51).

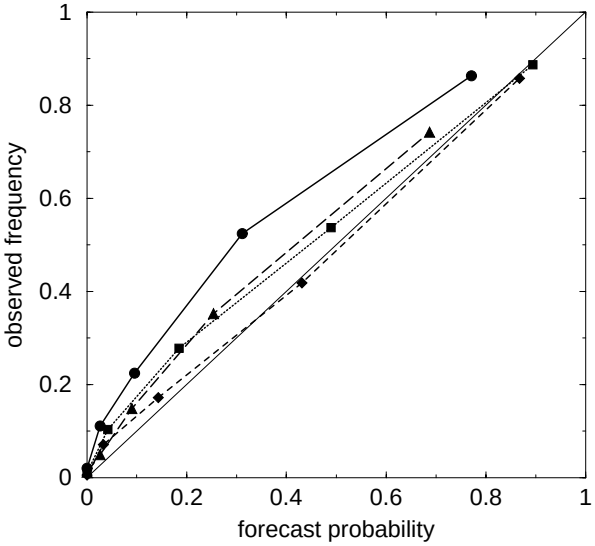


Figure 4. Same as Fig. 2, but the four winter seasons are considered separately: 1996-1997 (solid line, circles), 1997-1998 (dotted line, squares), 1998-1999 (short dashed line, diamonds), 1999-2000 (long dashed line, triangles). All grid points are considered together.

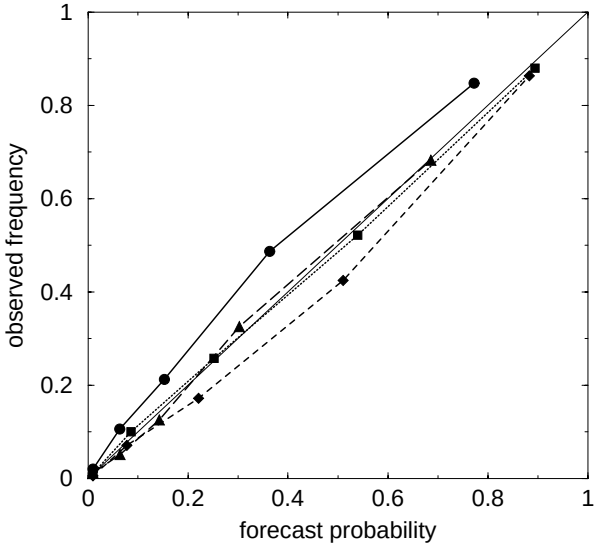


Figure 5. Same as Fig. 4, after calibration. As in Fig. 2, the four winter seasons have been considered together for computing calibration statistics. All grid points are considered together.

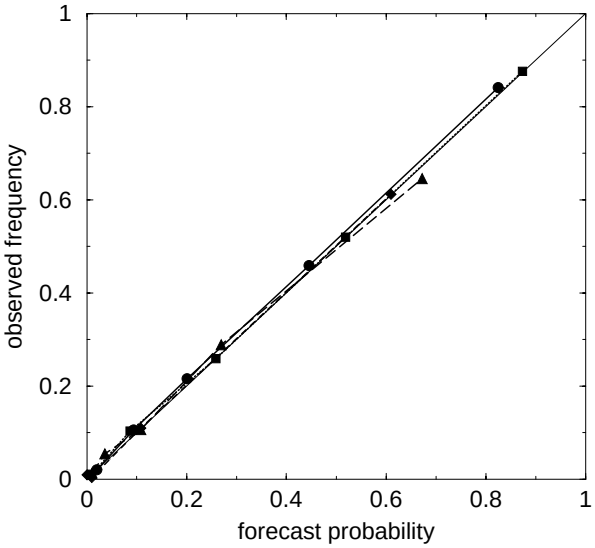


Figure 6. Same as Fig. 5, but the four winter seasons have been considered separately for computing calibration statistics. All grid points are considered together.

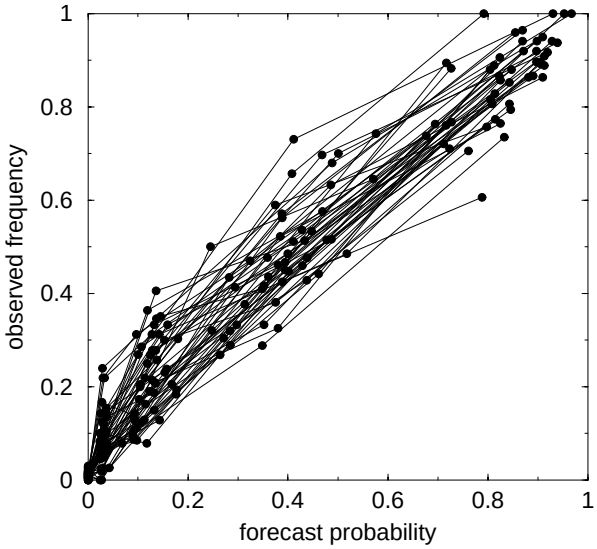


Figure 7. Same as the main graph of Fig. 2, before calibration, but the 48 grid points have been considered separately. The four winter seasons have been considered together.

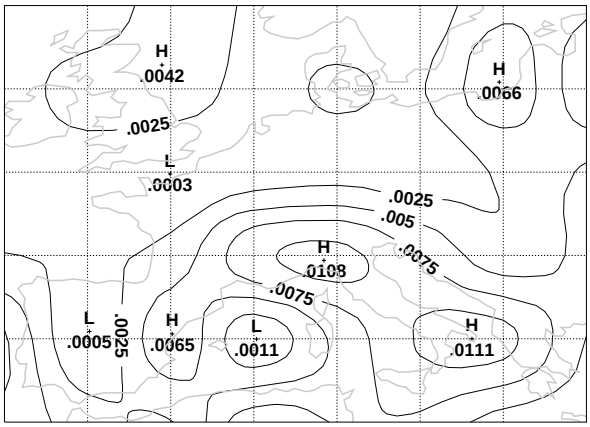


Figure 8. Reliability component of the Brier score computed at every grid point. Interval is 2.5 10^{-3}. Spatial average is 3.9 10^{-3}, standard deviation is 2.9 10^{-3}. The four winter seasons have been considered together.

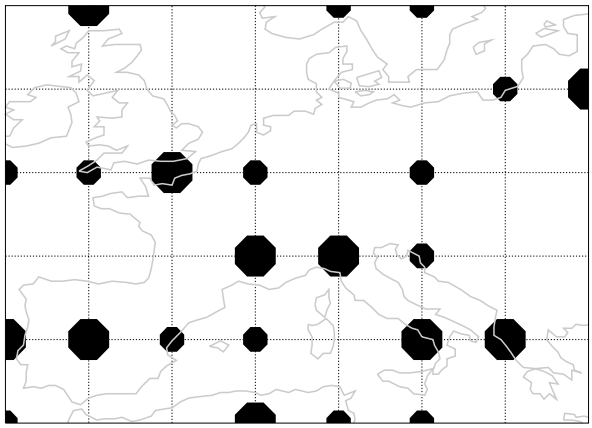


Figure 9. Significance level of the spatial variability of the reliability component of the Brier score, computed at every grid point, estimated from an empirical distribution of 1000 random computations (see text). Large symbols: significance above .95. Small symbols: significance above .90.

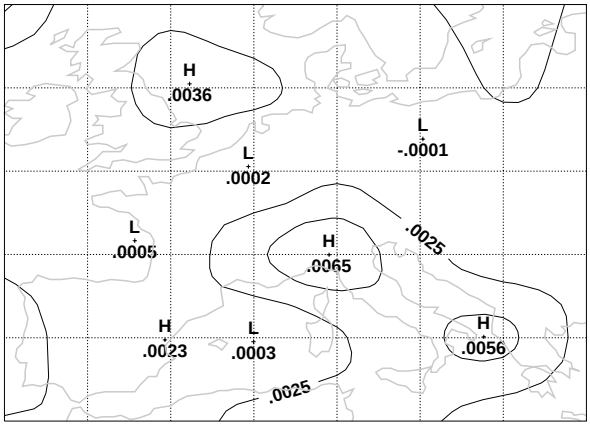


Figure 10a. Same as Fig. 8, after domain calibration. All grid points have been considered together for computing calibration statistics. The four winter seasons have been considered together. Spatial average is 2.0 10^{-3}, standard deviation is 1.5 10^{-3}.

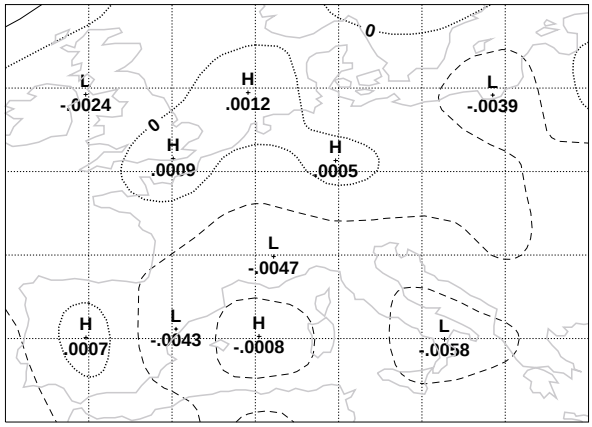


Figure 10b. Difference between the reliability component of the Brier score computed at every grid point after and before domain calibration (i.e. Fig. 10a - Fig. 8). Negative values indicate an improvement (dashed isolines), positive values indicate a deterioration (solid isolines). The dotted isoline indicates a null effect. Interval is 2.5 10^{-3}.

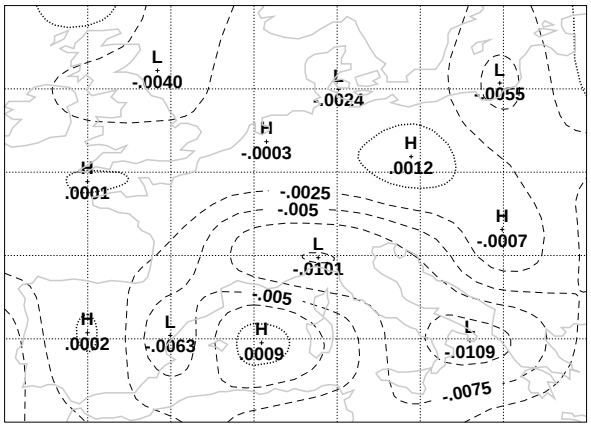


Figure 11. Same as Fig. 10b, but after and before local calibration. The four winter seasons have been considered together, for computing calibration statistics as well as for the evaluation.

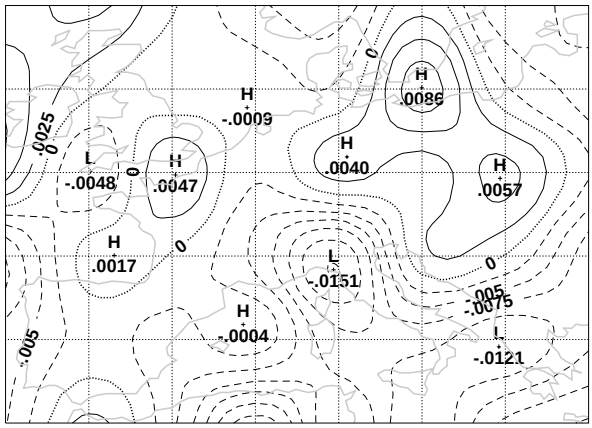


Figure 12. Same as Fig. 10b, but difference between the average reliability computed separately for the four winter seasons, after and before local calibration. The four winter seasons have been considered separately for computing calibration statistics.

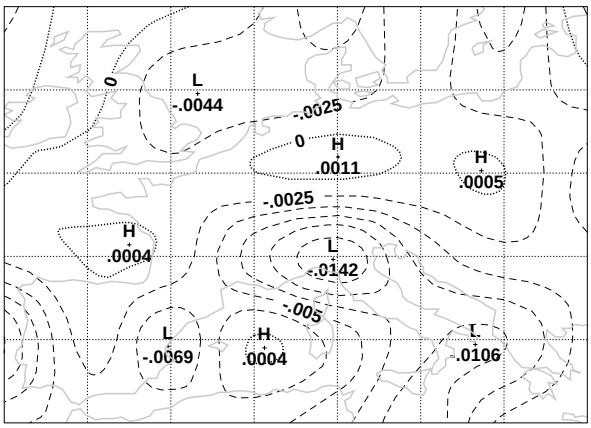


Figure 13. Same as Fig. 12, but the four winter seasons have been considered together for computing calibration statistics.

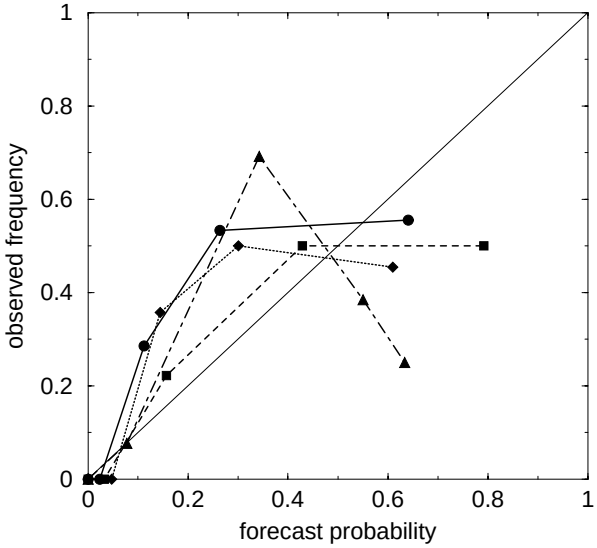


Figure 14. Same as Fig. 2, but for grid point 16 and winter season 1999-2000 only. Before calibration (solid, circles), after winterwise local calibration (dot-dashed, triangles), after local calibration based on the four winter seasons together (dotted, diamonds), after winterwise, local bias correction (dashed, squares). The reliability component of the Brier score is respectively 17.5 10^{-3}, 34.7 10^{-3}, 16.5 10^{-3} and 11.4 10^{-3}.

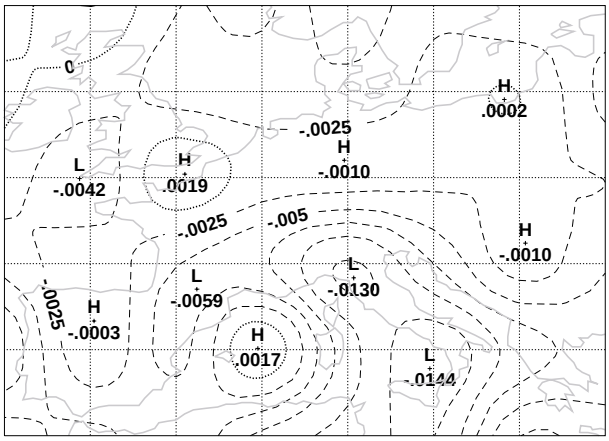


Figure 15. Same as Fig. 12, but after and before winter-wise local bias correction.

Chapitre 4

Estimation de la fiabilité de prévisions probabilistes à partir d'échantillons de taille réduite

Autour de l'article : "Estimation of the expected reliability of ensemble based probabilistic forecasts" (Atger, 2003)

1. Introduction

La décomposition du score de Brier a été introduite au chapitre 1. Les problèmes que pose l'estimation du terme de fiabilité ont été abordés au chapitre 3 à travers les conséquences de la variabilité spatiale et temporelle de la fiabilité. La question de la catégorisation des probabilités prévues, abordée brièvement au chapitre 3, est approfondie dans le présent chapitre. Il existe une catégorisation qu'on peut qualifier de naturelle, qui consiste à ne regrouper que des probabilités strictement égales, de $p=0/n$ à $p=n/n$ (si n est le nombre de membres de l'ensemble). Quand n est grand (par exemple : $n=51$ pour l'EPS du CEPMMT), cette catégorisation requiert des échantillons de taille conséquente afin que toutes les catégories soient représentées par un nombre suffisant de cas. Dans la plupart des évaluations les $n+1$ catégories naturelles sont regroupées de manière arbitraire, souvent sous la forme d'intervalles de 10% (par exemple : Mullen & Buizza, 2001). La catégorie [p=0] est parfois distinguée car elle est souvent la plus peuplée quand on considère des événements peu fréquents (c'est le cas par exemple pour le calcul des scores opérationnels du CEPMMT qui sont mis à la disposition des états membres).

En pratique, le nombre de cas regroupés dans des intervalles arbitraires varie beaucoup. En particulier, les faibles et/ou les fortes probabilités sont plus souvent représentées que les probabilités proches de la fréquence climatologique de l'événement considéré. Cette propriété, que j'ai proposé de traduire par *acuité* ('sharpness') est par ailleurs un des attributs de la qualité d'un ensemble. Elle

indique la variabilité des probabilités prévues. Un ensemble sans acuité prévoit toujours la même probabilité, il n'a donc aucune résolution (et ne présente aucun intérêt pratique, même s'il est fiable). Dans le cas d'un ensemble parfaitement fiable, l'acuité n'est rien d'autre que la résolution (qui indique la variabilité de la fréquence observée quand la probabilité prévue varie).

Du fait de l'acuité, et parce que les échantillons d'évaluation sont finis et relativement petits, certains intervalles sont susceptibles d'être insuffisamment peuplés. La première conséquence est que le niveau de correspondance entre la probabilité prévue dans ces intervalles et la fréquence observée correspondante ne sera pas significatif. La seconde conséquence est que le terme de fiabilité du score de Brier, calculé à partir de tous les intervalles, sera surévalué par rapport à ce qu'il serait si les intervalles étaient suffisamment peuplés. Le terme de fiabilité est en effet la moyenne quadratique des biais de chaque catégorie de probabilité, il est par conséquent systématiquement surévalué dans le cas où le biais estimé pour certaines catégories s'éloigne de zéro du fait d'une représentativité insuffisante de l'échantillon utilisé. Cet effet est facilement observable sur un diagramme de fiabilité obtenu à partir d'un échantillon trop petit et/ou à partir d'un trop grand nombre de catégories de probabilité (par exemple : Talagrand et al., 1997) : le terme de fiabilité du score de Brier étant la distance entre la courbe de fiabilité et la diagonale, tout "accident" de la courbe dû au sous-échantillonnage conduit à une surévaluation numérique (voir aussi la Fig. 1b de l'article présenté).

Le premier objectif du travail présenté dans ce chapitre consiste en une quantification de cette surévaluation. Le second objectif est de proposer une méthode d'évaluation de la fiabilité qui ne soit pas trop sensible à la petitesse des échantillons disponibles.

D'une façon générale, la vérification des prévisions météorologiques poursuit deux objectifs complémentaires. Le premier objectif, qui est celui d'un contrôle qu'on peut qualifier d'administratif, est d'évaluer *a posteriori* la qualité des

prévisions fournies aux utilisateurs. C'est le rôle, par exemple, des évaluations qui sont effectuées régulièrement par les services de Météo-France pour obtenir les quelques indicateurs de qualité qui seront fournis à la presse et serviront de support à la communication institutionnelle de l'établissement public. Ce type de vérification permet d'établir la qualité des prévisions extraites d'un échantillon de cas prévus dans le passé, mais il n'existe aucune raison de penser que le résultat obtenu soit généralisable à d'autres échantillons. En particulier, ce type de vérification ne vise pas à *prévoir* ce que sera la qualité des prévisions dans l'avenir. Il existe un second objectif de la vérification, qui est celui d'un contrôle qu'on peut qualifier de scientifique, consistant à estimer la qualité des prévisions en général, et pas seulement la qualité des prévisions qui constituent l'échantillon considéré. Dans ce cas on estime *a priori* la qualité des prévisions futures à partir de la qualité des prévisions passées. Plus précisément, on estime, à partir d'un échantillon de taille finie, la qualité des prévisions formant une population théorique de taille infinie. Une telle évaluation passe par l'utilisation d'un échantillon qui soit effectivement représentatif de cette population.

Le titre de l'article présenté dans ce chapitre précise qu'il s'agit d'estimer la fiabilité espérée ('expected reliability'). Il s'agit donc d'une évaluation a priori. Une évaluation a posterioreri fait nécessairement appel à une catégorisation des probabilités prévues qui reflète la variété des valeurs de probabilité qui sont effectivement fournies aux utilisateurs. En pratique, il est rare que les probabilités issues d'un ensemble soient arrondies avant d'être diffusées, par conséquent seule la catégorisation naturelle est acceptable. En revanche, dans le cas d'une évaluation a priori, toute catégorisation est acceptable dès lors qu'elle permet une évaluation crédible de la fiabilité à partir d'un échantillon de taille réduite. C'est à dire une catégorisation qui permet une estimation dont on n'ait aucune raison de penser qu'elle ne reflète pas fidèlement la fiabilité d'une population de prévisions qu'on pourrait constituer pendant une très longue période.

L'impact de la catégorisation des probabilités prévues sur le terme de fiabilité du score de Brier est d'abord évalué dans l'article. On propose ensuite d'utiliser un test statistique non paramétrique pour construire les catégories de manière optimale.

On peut aussi envisager d'évaluer la fiabilité sans passer par une catégorisation spécifique des probabilités prévues. La décomposition du score de Brier, sous sa forme continue, fait apparaître les fonctions $g(p)$ et $o(p)$ qui décrivent respectivement la fréquence de prévision de la probabilité p et la fréquence d'observation de l'événement quand la probabilité p est prévue. Une estimation de ces fonctions, à partir de l'échantillon disponible, est nécesaire et suffisante pour le calcul des termes de la décomposition du score de Brier. Cette estimation se fait généralement à l'aide d'une catégorisation et conduit à une forme discrète des fonctions. Rien n'empêche cependant, dans une perspective d'évaluation a priori, d'estimer ces fonctions par d'autres moyens. Il existe des exemples d'ajustement de la courbe de fiabilité par des fonctions paramétrées (Déqué, 2001) qui peuvent conduire à l'estimation de $o(p)$ et $g(p)$. Le choix du modèle statistique est cependant malaisé, dans la mesure où il n'y a aucune garantie pour que la fonction retenue, qui s'ajuste bien au jeu de données A, s'adapte aussi bien au jeu de données B. L'expérience montre en effet que les courbes de fiabilité n'appartiennent pas toutes à une même famille.

Dans ce chapitre on a utilisé les propriétés de la courbe ROC (outil de validation présenté dans le chapitre 2) pour estimer $o(p)$ et $g(p)$ par un ajustement linéaire des transformées gaussiennes des taux de détection et de fausse alarme. Contrairement à ce qui vient d'être mentionné pour les courbes de fiabilité, les courbes ROC qu'on rencontre dans diverses études ont en effet une allure caractéristique et semblent toutes appartenir, en pratique sinon en théorie, à une même famille. Cette estimation permet de reconstituer une courbe de fiabilité qui n'est presque plus affectée par le sous-échantillonnage, permettant ainsi une évaluation réaliste du terme de fiabilité du score de Brier. Cet ajustement permet

également d'évaluer la fiabilité des probabilités qui sont rarement prévues, par example les probabilitiés élevées qu'un événement rare se produise.

L'efficacité d'une procédure de calibration dépend beaucoup de la méthode utilisée pour évaluer la fiabilité (c'est une question qui a été discutée longuement dans le chapitre 3). Travaillant le plus souvent avec des échantillons de taille réduite, l'apprentissage nécessaire à la calibration (comme à toute correction statistique) est généralement peu satisfaisant et conduit à une amélioration relativement limitée de la fiabilité. Un ajustement de la courbe de fiabilité, par la méthode mentionnée ci-dessus, permet de connaître pour chaque probabilité prévue la fréquence observée espérée. Si cet ajustement est fiable, alors il doit conduire à une efficacité optimale de la procédure de calibration.

2. Données et définitions (section 2)

Dans cette section on rappelle seulement la définition du score de Brier et des termes de la décomposition proposée par Murphy (1973). Les données sont les mêmes que celles utilisées dans le chapitre 3.

3. Impact de la catégorisation (section 3)

On examine dans cette section les variations du terme de fiabilité du score de Brier en fonction du nombre de catégories qui sont utilisées pour la décomposition, lorsque les données considérées sont en nombre insuffisant pour échantillonner convenablement toutes les catégories naturelles (ce qui est généralement le cas en pratique, en particulier quand on travaille avec un ensemble de 51 membres comme l'EPS du CEPMMT). Le principal résultat est que le terme de fiabilité croît continument avec le nombre de catégories (Fig. 2a). En l'absence de plateau indiquant un nombre "pertinent" de catégories, seule une inspection visuelle du diagramme de fiabilité permet d'apprécier subjectivement la "vraie" correspondance entre probabilité prévue et fréquence observée, telle qu'on pourrait l'observer à partir d'un échantillon de très grande taille (Fig. 1).

4. Utilisation d'un test statistique (section 4)

Pour effectuer une catégorisation pertinente, il est assez naturel de chercher à constituer des catégories qui soient significativement différentes entre elles du point de vue d'un test statistique, c'est à dire pour lesquelles les fréquences observées ne peuvent être considérées comme indiscernables. Un test non paramétrique a été mis au point pour cette étude (voir en annexe). Il est basé sur les principes de ré-échantillonnage proposés par Hamill (1999) auxquels on a déjà eu recours aux chapitres 2 et 3. En pratique, pour obtenir des courbes de fiabilité « bien lisses », il faut une probabilité supérieure à 99% que les différences entre les fréquences observées ne soient pas le fait du hasard de l'échantillonnage (Fig. 3). Le nombre de catégories ainsi obtenues décroît évidemment avec la fréquence climatologique d'occurrence de l'événement considéré. Par conséquent l'efficacité de cette méthode est limité aux probabilités prévues assez fréquemment. La fiabilité des fortes probabilités d'un événement rare, par exemple, ne peut être évaluée.

5. Estimation par ajustement de la courbe ROC (section 5)

La méthode proposée dans cette section ne fait aucune hypothèse explicite sur la forme de la courbe de fiabilité. Il s'agit de tirer parti du fait que la courbe ROC (cf. chapitre 2) est ajustable par une droite après transformation de ses coordonnées en leurs transformées gaussiennes (Harvey et al., 1992). L'application de ce modèle repose sur l'hypothèse binormale : les distributions "signal" et "bruit" (c'est à dire, au sens de la théorie du signal, les distributions d'une variable de décision suivant que l'événement se produit ou ne se produit pas) sont supposées gaussiennes.

La validité de ce modèle est longuement discutée dans l'article. En pratique, l'ajustement est très bon dans tous les cas examinés (corrélation supérieure à 99%) (Fig. 5) ce qui confirme les résultats obtenus par Richardson (2000) avec des données différentes provenant également de l'EPS du CEPMMT. L'ajustement des courbes de fiabilité est visuellement très réussi, et proche de ce qu'on aurait

pu tracer subjectivement (Fig. 6). La validité de l'ajustement est mise en évidence objectivement par la reconstitution de la partie basse de la courbe de fiabilité (très bien échantillonnée) à partir d'une fraction d'environ 10% des cas de l'échantillon considéré (Fig. 7). Ces résultats semblent indiquer qu'il existe effectivement une variable de décision sous-jacente, obtenue par transformation monotone de la probabilité prévue, qui satisfait à l'hypothèse bi-normale.

L'ajustement de la courbe ROC permet également d'évaluer la fiabilité des probabilités élevées d'événement rare (Fig. 8). Ces probabilités sont rarement prévues, mais sont pourtant perçues par la plupart des utilisateurs comme les plus importantes, celles pour lesquelles la fiabilité est un enjeu, tandis que les faibles probabilités sont souvent ignorées (Atger, 2001). L'ajustement peut conduire dans ce cas à une forme d'extrapolation, dont la validité est évidemment limitée par la petitesse de l'échantillon considéré. On constate cependant que des données correspondant à une courbe de fiabilité très bruitée permettent un très bon ajustement de la courbe ROC correspondante.

Le terme de fiabilité du score de Brier, estimé par ajustement de la courbe ROC, ne subit pas de sur-évaluation numérique quand on réduit, jusqu'à un certain point, le nombre de cas considérés (Fig. 9).

6. Conséquences pour la calibration (section 6)

Comme dans le chapitre 3, on utilise la méthode "classique" de calibration, qui consiste simplement à prévoir comme probabilité calibrée la fréquence observée dans un échantillon d'apprentissage, différent de celui sur lequel on effectue l'évaluation, lorsque le même nombre de membres de l'ensemble prévoient l'événement considéré (Zhu et al., 1996). En raison de la petitesse des échantillons considérés, tant pour l'apprentissage que pour l'évaluation, la calibration ne permet pas d'améliorer la fiabilité des probabilités d'un événement rare tel qu'une anomalie supérieure à 2 écarts types (Fig. 11). L'utilisation de la méthode d'ajustement décrite dans la section 5 permet de prévoir comme probabilité calibrée, non plus la fréquence observée, mais la fréquence *espérée*,

telle qu'elle peut être estimée à partir de l'échantillon d'apprentissage. On montre que cette méthode permet d'obtenir une amélioration de la fiabilité, y compris dans le cas d'un événement rare (Fig. 12 et 13).

7. Remarques conclusives

Les raisons exactes pour lesquelles l'ajustement de la courbe ROC est aussi efficace restent peu claires à l'issue de ce travail. Un argument avancé pour rendre compte de la robustesse apparemment universelle du modèle binormal est la limitation du nombre de catégories considérées pour construire la courbe ROC (Hanley, 1988). Cet argument n'est pas valide dans le cas de prévisions issues d'un ensemble de 51 membres. D'autres auteurs invoquent le théorème central limite (Green & Swets, 1966), qui pourrait s'appliquer dans le cas d'une prévision issue d'un ensemble. L'événement météorologique qu'on cherche à prévoir peut en effet être considéré, jusqu'à un certain point, comme la résultante d'un grand nombre de causes naturelles indépendantes. Par conséquent la variable de décision qui conditionne l'occurrence de cet événement pourrait être la somme d'un grand nombre de variables indépendantes. Cet argument est supporté par le fait que des simulations effectuées à partir d'ensembles Gaussiens ne permettent pas toujours d'obtenir un ajustement de la courbe ROC qui soit aussi réussi que celui observé dans la présente étude à partir de l'EPS (expériences conduites récemment, et dont le résultat demande confirmation). La forme des fonctions o(p) et g(p), facilement reproduite par la simulation, semble ne pas être en cause. La principale différence est que l'occurrence de l'événement à prévoir est entièrement conditionnée par une variable aléatoire unique, et non par un ensemble de causes météorologiques réelles.

Annexe

Description de la méthode utilisée pour établir une catégorisation des probabilités prévues (section 4)

L'objectif de la méthode est d'évaluer dans quelle mesure deux catégories de probabilité prévue doivent être regroupées en une seule et même catégorie, au motif que la fréquence observée de l'événement considéré est la même avant et après regroupement. Comme dans les chapitres 2 et 3 (annexes) la méthode est basée sur un test statistique non paramétrique qui requiert un grand nombre de tirages aléatoires effectués dans l'échantillon disponible.

Considérons les M_i cas pour lesquels la probabilité prévue est $p_i=i/N$ (c'est à dire lorsque i membres de l'ensemble parmi N prévoient l'événement considéré) et les M_j cas pour lesquels la probabilité prévue est $p_j=j/N$ ($j>i$). $o(p_i)$ (resp. $o(p_j)$) est la fréquence d'observation de l'événement dans les cas où la probabilité prévue est i/N (resp. j/N). On définit o_1 comme la fréquence d'observation de l'événement dans les M_i+M_j cas pour lesquels la probabilité prévue est i/N ou j/N. On définit o_2 comme la fréquence d'observation de l'événement dans M_i cas extraits par tirage aléatoire parmi les M_i+M_j cas pour lesquels la probabilité prévue est i/N ou j/N. La procédure consiste à répéter 1000 fois le tirage aléatoire et le calcul de o_2 afin de construire une distribution empirique de la différence absolue $|o_1-o_2|$. En comparant la différence absolue $|o(p_i)-o(p_j)|$ à cette distribution, on déduit la probabilité (empirique) que la variation de la fréquence d'observation soit significative quand on augmente de i à j le nombre de membres de l'ensemble prévoyant l'événement.

Comme dans le chapitre 3, l'hypothèse d'indépendance nécessaire à l'application du test statistique est supposée remplie. On devrait cependant, en toute rigueur, corriger le résultat pour tenir compte de l'existence de corrélations spatiales et temporelles.

La catégorisation proprement dite s'effectue par un processus itératif. On commence par tester les différences entre la catégorie p_0 (aucun membre de l'ensemble ne prévoit l'événement) et les catégories $p_1 \ldots p_j$, jusqu'à trouver une différence significative pour p_j. On teste alors la différence entre la catégorie p_j et les catégories p_k ($k>j$) jusqu'à trouver une différence significative. On répète l'opération jusqu'à épuisement et on constitue la nouvelle catégorisation en regroupant les catégories non significativement différentes.

Bibliographie

Atger, F., 2001: Performance and usefulness of ensemble based probabilistic forecasts. 8th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF (Proceedings, 29-31).

Atger, F., 2003: Estimation of the expected reliability of ensemble based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, soumis.

Déqué, M., 2001: Seasonal predictability of tropical rainfall: probabilistic formulation and validation. *Tellus*, **53A**, 500-512.

Green, D. M. & J. A. Swets, 1966: Signal detection theory and psychophysics. New York, Wiley and Sons.

Hamill, T. M, 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

Hanley, J. A., 1988: The robustness of the binormal assumptions used in fitting ROC curves. *Med. Decis. Making*, **8**, 197-203.

Harvey, L.O., K.R. Hammond, C.M. Lusk & E.F. Mross, 1992. The application of signal detection theory to weather forecasting behaviour. *Mon. Wea. Rev.*, **120**, 863-883.

Mullen, S. L. and R. Buizza, 2001. Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-661.

Richardson, D. S, 2000. Skill and economic value of the ECMWF Ensemble Prediction System, *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings, Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts, 1-25.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conference on Weather Analysis and Forecasting, Norfolk, Virginia, Amer. Meteor. Soc., J79-J82.

Estimation of the expected reliability of ensemble-based probabilistic forecasts

Frédéric Atger
Météo-France (Toulouse, France)

Submitted to the Quarterly Journal of the Royal Meteorological Society

31 January 2003

Frédéric Atger, Météo-France (DPREVI/COMPAS)
42, Av. G. Coriolis - 31057 Toulouse cedex - France
Tph: 05 61 07 85 37 - Fax: 05 61 07 84 53
E-mail: frederic.atger@meteo.fr

Summary

Reliability is an essential attribute of the quality of probabilistic forecasts. It is traditionally estimated by defining a number of arbitrary probability categories. Due to the relatively small size of verification samples, reliability is likely to be unrealistically estimated, at least for certain categories of forecast probability, especially in the case of forecasting rare events. Significance tests are used in this study in order to determine an appropriate categorisation of forecast probabilities for the estimation of reliability. For events occurring frequently this method leads to credible estimates of the performance for the whole range of forecast probabilities. On the other hand, the reliability of higher probabilities of infrequent events cannot be estimated with confidence. A parameterisation scheme has been designed for estimating reliability from limited samples, even in the case of rare events and higher probabilities. The procedure consists in fitting a relative operating characteristic (ROC) curve under the bi-normal assumption. The validity of the method is discussed by testing its ability to estimate reliability from truncated verification samples. The positive impact of a basic method of calibration is increased when it is applied after an estimation of reliability through a fitting of the ROC curve.

KEYWORDS: Ensemble prediction Brier score ROC curve Bi-normal model Calibration

1. Introduction

The validation of ensemble prediction systems (EPS) is mainly based on the verification of ensemble-based probabilistic forecasts. Since December 1992, both the U.S. National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-range Weather Forecasts (ECMWF) have been producing operational forecasts based on ensemble prediction (Tracton and Kalnay 1993, Palmer et al. 1993). Several other centres worldwide have since implemented operational or quasi operational ensembles (e.g. Houtekamer et al. 1996, Kobayashi et al. 1996). Proper validation requires quantitative elements of comparison between the different strategies that can be followed in the development of an operational EPS: model characteristics, number of ensemble members, method for generating perturbations, etc. (Atger 1999).

In a broad sense, forecast verification consists in a comparison of a collection of M forecasts f_i to a collection of M verifying observations x_i. The level of correspondence indicates how accurate the forecasting system is (Murphy and Winkler 1987). In the case of an EPS, every forecast f_i is an ensemble of N members, supposed to be independent realizations of the probability density function (pdf) of the future state of the atmosphere. The forecast f_i is thus an empirical estimate of this pdf. On the other hand, x_i is still a single observation, reflecting the fact that a perfect probabilistic forecast is a deterministic forecast that proves correct. Therefore, ensemble verification consists in a comparison of a collection of M pdf estimates to a collection of M observations. Because a pdf estimate and a single observation have not the same statistical nature, different methods can be contemplated for this comparison. A direct comparison is possible when considering the single verifying observation as an empirical estimate of a pdf represented by a Dirac function. This approach is followed when computing the Continuous Ranked Probability Score, that can be defined as the integrated squared difference between the predicted and observed cumulative distributions (Hersbach 2000).

A different approach consists in considering together the cases when the estimated pdf is the same, or at least similar, and comparing this pdf estimate to the corresponding distribution of observations. A variety of criteria can be considered for categorizing similar pdf estimates and evaluating the closeness between the forecast and observed distributions. Most of the time, only certain aspects of the distributions are considered. Ensemble distributions are often categorized according to the number of ensemble members forecasting a specified

meteorological event. This leads to the comparison between a forecast probability p, defined as the proportion of ensemble members forecasting an event, and the observed frequency of this event when p is forecast. This comparison is generally quantified by the mean square difference between the forecast probability and the corresponding observed frequency, i.e. the reliability term of the Brier score (Brier 1950, Murphy 1973). More generally, reliability indicates to which extent a pdf estimate proves close, a posteriori, to the distribution of observations when this pdf estimate is predicted. Besides reliability, resolution is the second essential aspect of the performance of a probabilistic forecast. Resolution is the variability of the observed frequency of an event, when the forecast probability of this event varies. In a more general sense, resolution indicates the variability of the distribution of observations, when the predicted pdf estimate varies.

Although ensemble distributions leading to the same probability p are likely to exhibit large differences with respect to other aspects, the verification of ensemble-based probabilistic forecasts of some selected meteorological events is by far the most common approach for validating ensemble prediction systems (e.g. Mullen and Buizza 2001). This is obviously due to the fact that most operational products derived from ensemble prediction systems are probabilistic forecasts. The verification of ensemble-based probabilistic forecasts has thus two purposes: validating the system and evaluating the performance of end products.

As noticed by Murphy (1993), the lack of reliability can be seen as a probability bias. Therefore, several authors have proposed a calibration of probabilistic forecasts in order to improve their reliability. Calibration consists in a statistical correction of the forecast probability, based on previous verification. For example, in the case of the probabilistic forecast of a dichotomous event, if the observed frequency in the past was 30% when the forecast probability was 40%, the calibrated probability will be 30% when the raw probability is 40% (Zhu et al. 1996). Alternatively, a statistical correction can be applied to the ensemble distribution, in order to get the pdf estimate closer to the observed distribution of observations (Hamill and Colucci 1998).

Verification is basically an evaluation of the performance of a finite sample of forecasts. Strictly speaking, the result of this *a posteriori* evaluation is only valid for the considered sample. There is no reason for generalizing the result of the verification to other forecast samples. On the other hand the considered sample of forecasts may be representative, to a certain extent, of the population of all possible forecasts. The result of this *a priori* estimation

may thus indicate, to a certain extent, the expected performance of the forecasting system. An example of a posteriori evaluation is given by the forecast scores used by national weather services or private weather companies for public information and promotion campaigns. These scores, unless explicitly stated, do not indicate the expected forecast performance. In the present study, as in most research works, verification rather aims at an a priori estimation of the intrinsic, expected performance of the forecasting system, independent, to a certain extent, of the considered sample.

Estimating reliability and resolution requires a categorisation of probabilistic forecasts. This categorisation is needed for calibration too. In the case of an a posteriori evaluation, probability categories should reflect the variety of probabilistic forecasts that have been considered for verification. Probabilities computed from an EPS distribution take N+1 different values, from 0/N to N/N (if N is the number of ensemble members), but end probabilistic forecasts may take only a few different values, e.g. from 0 to 1 with a 0.1 increment, depending on the users requirements. In the case of an a priori estimation, categorisation is arbitrary. Categories should be designed with the aim of getting results that give a true picture of the performance, i.e. results that can be generalized. In particular, significant results are more likely obtained when probability categories are designed in order to evenly distribute the cases among them, at least approximately, in order to avoid under-sampling. Another strategy might consist in a parameterisation of the distribution of forecast probabilities and observed frequencies.

The aim of the work described in the present article is to evaluate different methods for the estimation of the reliability of ensemble-based probabilistic forecasts. Consequences for calibration are investigated too. The paper is organized as follows. Definitions are given in section 2, as well as a description of the data. The effect of the number of probability categories on the estimation of the reliability and resolution terms of the Brier score is investigated in section 3. Significance tests are used in section 4 for defining a proper categorisation for estimating reliability. A method for estimating reliability through parameterisation is described in section 5. Consequences for the calibration of probabilistic forecasts are discussed in section 6. Results are summarized in section 7.

2. Data and definitions

(a) Data

Four consecutive winter seasons (December to February) have been considered, from 10

December 1996 to 28 February 2000 (351 days). During these four years the operational version of the ECMWF ensemble prediction system has been improved several times, as well as the model on which it is based, but the horizontal resolution remained the same (T_L159), as well as the number of ensembles members (N=50+1). For this reason one can assume a certain stability of the performance.

Forecasts considered in the study are +96 h EPS-based probabilities of 850-hPa temperature anomaly at 48 grid points over Europe (35N, 60N, 10W, 25E, 5x5° grid). The sample size is M=351x48=16848 cases, obviously not independent because of space and time correlations. For the sake of simplicity these correlations will be disregarded, and the 16848 cases will be considered as independent realizations of a single random variable.

The verification reference is the ECMWF high-resolution analysis that was operational during the considered period (T_L319), interpolated at every grid point. Given the high density of upper air observations over the considered area (radio-sondes, aircraft observations, etc.), these local interpolations are assumed to reflect the true state of the atmosphere.

The climate reference, for the definition of anomalies, has been computed from the ECMWF 15-year reanalysis (Gibson et al. 1997).

(b) Brier Score

The Brier score (BS) is defined for a dichotomous event as the mean square error of the probability forecast:

$$BS = \frac{1}{M}\sum_{i=1}^{M}(p_i - o_i)^2 \qquad (1)$$

where M is the number of considered cases, p_i is the forecast probability, o_i is the verifying observation (1 if the event occurs, 0 if it does not).

Considering a continuous range of forecast probabilities, from 0 to 1, the Brier score can be rewritten as:

$$BS = \int_0^1 ([1-o(p)]p^2 + o(p)[1-p]^2)g(p)dp \qquad (2)$$

where $g(p)$ is the frequency with which the probability p is forecast and $o(p)$ is the frequency with which the event is observed when the probability p is forecast (Talagrand et al. 1997).

(c) Reliability and resolution

Under the form given in Eq. (2), the Brier score is easily transformed into the decomposition initially proposed by Murphy (1973):

$$BS = \int_0^1 [p - o(p)]^2 g(p) dp - \int_0^1 [o - o(p)]^2 g(p) dp + o(1-o) \qquad (3)$$

where o is the overall frequency with which the event is observed (Talagrand et al. 1997). The first part of the decomposition is called the reliability term. It is the integration, over the whole range of forecast probabilities, of the square difference between the probability and the observed frequency of the event. The second part of the decomposition is called the resolution term. It is the variance, over the whole range of forecast probabilities, of the observed frequency of the event. The third part is called the uncertainty term. It is the variance of the observations, which does not depend on the forecast system and reflects the intrinsic difficulty in forecasting the observations.

3. Effect of the categorisation

Estimates of the functions *g(p)* and *o(p)* are needed for the computation of the reliability and resolution terms of the Brier score from Eq. (3). In the field of ensemble verification, authors generally consider discrete forms of *g(p)* and *o(p)*, either with N+1 forecast probabilities from 0/N to N/N (Atger 1999), or with a limited number of probability categories, for instance 0.1 intervals (Mullen and Buizza 2001). The former is the only admissible choice if an *a posteriori* evaluation is required, unless end forecasts are effectively issued under the form of a limited number of probability categories (which is not usually the case). For an *a priori* evaluation, as in the present study, any estimation of *g(p)* and *o(p)* from the actual distributions of forecasts and verifications is acceptable. This estimation is traditionally done by categorizing forecast probabilities into a certain number of classes, the common choice of 0.1 intervals being just as arbitrary as any other choice. Another approach for estimating *g(p)* and *o(p)* consists in a parameterisation, as proposed in section 5.

Generally the number of verification cases in each category varies significantly, due to the (desirable) sharpness of the forecast distribution: higher and/or lower probabilities are more often forecast than probabilities lying close to the climate frequency. Some categories are thus more likely to suffer from a lack of data. In order to get significant results and avoid under-sampling effects, an adequate categorisation may consist in evenly distributing the cases, as

far as possible, among the different categories. This has been done for n categories, n varying from 1 (all probabilities considered together) to N+1 (probabilities from 0/N to N/N). Note that when $n<<N$ the verification cases are indeed evenly distributed among the categories, but this is obviously not the case when $n \approx N+1$. An example of reliability curve is shown in Fig. 1 for an anomaly above 1 standard deviation, for n=10 (panel a) and n=52 (panel b). The curve indicates the correspondence between the forecast probability (averaged into the category) and observed frequencies. The number of cases that have been considered for drawing the different points of the curve shown in the panel (b) is obviously too small to get a significant picture of the performance, while it does not seem to be the case for the curve shown in panel (a).

Graphically, the reliability term of the Brier score is the weighted, squared distance between the reliability curve and the 45° line. This term is likely to be numerically overestimated when it is computed from an insufficient number of cases, leading to a noisy reliability curve as the one shown in Fig. 1b. Similarly, the resolution term of the Brier score is the weighted, squared distance between the reliability curve and the horizontal line indicating the sample frequency of the event. The lack of sampling thus leads to a numerical over-estimation, generally small compared to that of the reliability term because the 45° line is closer to the reliability curve than the horizontal line indicating the sample frequency of the event. Figure 2 (top panel) shows the decrease of the reliability term of the Brier score with the number of probability categories. Apparently, there is no sign of a plateau that might indicate the better choice for a proper estimate of the performance: reliability increases steadily with the number of categories. On the contrary, resolution is hardly affected by the number of categories, except when this number is below 10 (Fig. 2, middle panel). Estimating resolution thus appears much easier than reliability. For this reason, this study will focus on reliability only in the following.

Note that the Brier score computed from Eq. (2) and Eq. (3) is numerically different from the discrete, traditional form defined in Eq. (1), unless the integration is approximated from all N+1 probability categories. The difference is negligible when the number of categories is sufficiently large (above 10). This is due to the fact that, in practice, the reliability term is generally much smaller than the resolution term, so that it contributes little to the Brier score. For a very small number of categories the resolution term is underestimated, so that the decomposition gives a wrong estimation of the Brier score (Fig. 2, bottom panel).

4. Defining a proper categorisation

It has been shown in the previous section that the reliability term of the Brier score increases monotonically with the number of probability categories used for its estimation. Practically, the level of noise exhibited by the reliability curve indicates to which extent sampling limitations have a detrimental impact on the estimation of reliability. Because a careful inspection of reliability diagrams is not always possible, there is a need for an objective method for computing reliable estimates of the reliability term of the Brier score from limited samples. One possibility consists in evenly distributing the verification cases into a limited number of probability categories, as mentioned in the previous section. However, there is some arbitrariness in the choice of the number of categories and/or the minimum number of cases that are considered together into the same category.

A possible strategy for determining a proper categorisation might be based on the fact that two categories should be merged if and only if the conditional probability of the event is not significantly different after merging than before. A resampling procedure has been designed to test the significance of this difference, i.e. to test to which extent the variations of the frequency of the considered event are significant when increasing the number of ensemble members forecasting this event. The method consists in the construction of an empirical distribution of differences that are not statistically significant (Hamill 1999).

Consider the categories i and j corresponding to forecast probabilities $p_i=i/N$ and $p_j=j/N$ respectively (i=0 to N, $j=i+1$ to N). $o(p_i)$ ($o(p_j)$) is the observed frequency of the event when considering the M_i (M_j) cases belonging to category i (j). o_1 is defined as the observed frequency of the event when considering the M_i+M_j cases belonging to either one or the other category. Similarly, o_2 is defined as the observed frequency of the event when considering M_i cases randomly extracted from the M_i+M_j cases belonging to either one or the other category. The procedure consists in computing o_2 1000 times, in order to generate an empirical distribution of the absolute difference $|o_1-o_2|$. The probability that the actual difference $|o(p_i)-o(p_j)|$ belongs to this distribution is then evaluated. It is an empirical estimate of the probability that the variation of the frequency of the event is not significant when increasing from i to j the number of ensemble members forecasting this event.

The process is iterative: starting from the category p_0 (no member forecasting the event), the difference with the categories p_i (i=1 to N), is tested until there is a significant difference for p_j. Significance testing is then applied to differences between the category p_j and the category

p_k (k=j+1 to N), until there is a significant difference. This is repeated until all categories have been formed.

Figure 3 shows the reliability diagram obtained through this procedure, for an anomaly above 1, 2 and 2.2 standard deviations (panels a, b and c, respectively). The significance threshold has been fixed to 0.99 (less than 1% probability that the difference before and after merging is not significant). Lower significance thresholds (e.g. 0.95) lead to rather noisy reliability curves (not shown). For an anomaly above 1 standard deviation the obtained categorisation (7 categories) is close to the one defined by 0.1 intervals that is traditionally used for evaluating probabilistic forecasts (panel a). The number of categories is much reduced when considering less frequent events: 4 categories for 2 standard deviations (panel b), 3 categories for 2.2 standard deviations (panel c). This means that it is not possible to estimate with high confidence the reliability of higher probabilities of rare events. Although the sample considered in this study consists of data that have been accumulated for 4 years, the reliability of forecast probabilities above 70% of an event occurring in 1% of the cases (anomaly above 2.2 standard deviations) cannot be estimated (panel c).

5. Estimation through parameterisation

(a) Fitting the ROC curve

As mentioned briefly in section 3, an estimation of $g(p)$ and $o(p)$ might be achieved through parameterisation. When considering a reliability diagram such as the one shown in Fig. 1 (panel b), one would like to fit the reliability curve in order to get rid of the noise obviously due to the limited size of the sample. This has been done for example by Déqué (2001) for reliability diagrams obtained in a seasonal prediction experiment. Yet the form of reliability curves is highly variable in practice and the choice of a parametric function is not straightforward. In this study it is proposed not to fit directly the reliability curve, but to take advantage of some properties of another verification tool, the Relative Operating Characteristic (ROC) curve.

The ROC curve is a plot of the Hit Rate (HR) i.e. the proportion of occurrences that have been successfully forecast, vs. the False Alarm Rate (FAR) i.e. the proportion of erroneously forecast occurrences. When evaluating a probabilistic forecast, $HR(p_k)$ and $FAR(p_k)$ are computed for K probability thresholds p_k by considering that the occurrence is forecast when $p \geq p_k$ (Mason 1982). The ROC curve thus consists of K points $(FAR(p_k),HR(p_k))$. In the case of ensemble-based probabilistic forecasts, it is natural to consider N+1 probability thresholds

of the form $p_k=k/N$ (k=0 to N). With the definitions given in section 2, $FAR(p_k)$ and $HR(p_k)$ can thus be written as integrals of $o(p)$ and $g(p)$:

$$HR(p_k)=\frac{1}{o}\int_{p_k}^{1} o(p)g(p)dp \qquad FAR(p_k)=\frac{1}{1-o}\int_{p_k}^{1}[1-o(p)]g(p)dp \qquad (4)$$

Another set of $K<N+1$ probability thresholds would be acceptable for computing *HR* and *FAR*, leading to a specific categorisation of predicted categories. However, because the computation of *FAR* and *HR* is cumulative (the number of considered cases increasing when the probability threshold p_k decreases) the "natural" categorisation appears much less sensitive to under-sampling effects than the categorisation from which the reliability curve is built. For this reason smooth ROC curves can be obtained through a computation of *HR* and *FAR* according to Eq. (4) even from rather small samples and in the case of rare events.

In the field of medicine and experimental psychology, it is common practice to fit the ROC curve with the so-called *bi-normal model*. The basic assumption is that there exists an underlying, generally unknown variable ξ, often called *decision variable*, defined for each verification case, and whose variations reflect the uncertainty related to the occurrence of the considered event. The *signal distribution* $f_s(\xi)$ is defined as the a posteriori distribution of ξ given the event occurs, while the *noise distribution* $f_n(\xi)$ is the a posteriori distribution of ξ given the event does not occur. With these definitions, *HR* and *FAR* are integrations of f_s and f_n above a critical value of the decision variable ξ_c:

$$HR(\xi_c)=\int_{\xi_c}^{\infty} f_s(x)dx \qquad FAR(\xi_c)=\int_{\xi_c}^{\infty} f_n(x)dx \qquad (5)$$

Under the bi-normal assumption, i.e. assuming f_s and f_n are normal, *HR* and *FAR* can both be expressed as integrations of the standard normal distribution f:

$$HR(\xi_c)=\int_{z_s(\xi_c)}^{\infty} f(x)dx \qquad FAR(\xi_c)=\int_{z_n(\xi_c)}^{\infty} f(x)dx \qquad (6)$$

with $z_s(\xi_c)=\frac{\xi_c-\mu_s}{\sigma_s}$ and $z_n(\xi_c)=\frac{\xi_c-\mu_n}{\sigma_n}$ where μ_s and σ_s (μ_n and σ_n) are respectively the mean and the standard deviation of f_s (f_n).

$z_{HR}=z_s(\xi_c)$ and $z_{FAR}=z_n(\xi_c)$ are often called the *standardized normal deviates* of *HR* and *FAR*

respectively, since $HR=F(z_{HR})$ and $FAR=F(z_{FAR})$ where $F(x)$ is the cumulative distribution function of the standardized normal random variable:

$$HR = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{HR}} e^{-\frac{x^2}{2}} dx \qquad FAR = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{FAR}} e^{-\frac{x^2}{2}} dx \qquad (7)$$

From their definition z_{HR} and z_{FAR} are linearly related:

$$z_{HR} = z_{FAR} \frac{\sigma_n}{\sigma_s} + \frac{\mu_n - \mu_s}{\sigma_s} \qquad (8)$$

Therefore, under the bi-normal assumption, an ROC curve is a straight line after transformation of its x and y coordinates into their standardized normal deviates according to Eq. (7). Note that the decision variable ξ may remain unknown, as well as the distributions f_s and f_n.

(b) Validity of the bi-normal model

In the case of probabilistic forecasts, it seems logical to consider the forecast probability as a decision variable. On the other hand, the (signal) distribution of the forecast probability given the event occurs $f_s(p) = \frac{1}{o} o(p) g(p)$ and the (noise) distribution of the forecast probability given the event does not occur $f_n(p) = \frac{1}{1-o}[1-o(p)]g(p)$ are generally not normal. When the considered event is not frequent, as in the present study, the noise distribution is very close to the unconditional distribution of the forecast probability. This distribution is far from normal (Fig. 4, bottom panel), as is the strongly U-shaped signal distribution that indicates the relatively poor ability of the forecast to discriminate between the occurrence and the non-occurrence of the event (Fig. 4, top panel).

In practice, however, the bi-normal model fits rather well the data. Figure 5 (panel a) shows an example of ROC curve for an anomaly above 1 standard deviation. The standardized normal deviates z_{HR} and z_{FAR} have been numerically determined according to Eq. (7) from a standard Gaussian distribution sampled with 10000 points from -5σ to +5σ. The linear correlation between z_{FAR} and z_{HR} is impressively high (Fig. 5, panel b). In all the cases that have been considered in this study the correlation coefficient was never below 0.99, even for rare events. This result confirms those obtained by Richardson (2000) for EPS based probabilistic forecasts of precipitation amounts and 850-hPa temperature anomaly. There

exists consistent empirical evidence from a number of studies that the bi-normal model allows an excellent fit of ROC curves built from probabilistic weather forecasts (Mason 1982). The bi-normal model has also proven highly efficient in various other experimental fields, especially in the medical area (Swets 1986). This apparent robustness of the bi-normal assumption is discussed further at the end of section 7.

In the present study, how is it possible to interpret (if not explain) the fact that the bi-normal model allows a good fit of the data? Although it is rather natural to take the forecast probability as a decision variable, any monotonic transformation of this probability leads to a potential decision variable. The result shown in Fig. 5 indicates that among these transformations, one leads to signal and noise distributions that are sufficiently close to normal for applying successfully the bi-normal model. In order to get some hint about what might be this transformation, with respect to the data that have been considered in this study, the forecast probabilities p and q for which $HR(p) = FAR(q)$ have been plotted one against the other (not shown). If the signal and noise distributions were both normal, it comes from Eq. (6) that there would be a linear relationship between p and q. Since f_S and f_n are obviously not normal (Fig. 4), it is not surprising that no linear relationship is found. However, the relationship is almost linear when p and q are plotted along log axes. When considering $\log(p)$ as a decision variable the distributions f_S and f_n look indeed closer to a Gaussian than it is the case in Fig. 4 (not shown).

(c) Estimation of o(p) *and* g(p)

The method proposed here for estimating $o(p)$ and $g(p)$ consists in 3 steps: (1) fitting the ROC curve under the bi-normal assumption, i.e. linear fitting after transformation of *HR* and *FAR* into z_{HR} and z_{FAR}; (2) estimating *HR* and *FAR* from the fitted ROC curve; (3) finally estimating *o(p)* and *g(p)* from *HR* and *FAR*.

The first step is described in the previous sub-section.

The second step consists in an orthogonal projection of the (*FAR*,*HR*) points of the original ROC curve onto the fitted curve, the latter being sampled with 1000 points (Fig. 5, panel a).

The third step in detailed in the present sub-section. After differentiation with respect to p, Eq. (4) becomes:

$$g(p) = -[o\frac{dHR}{dp} + (1-o)\frac{dFAR}{dp}] \qquad o(p) = -\frac{o}{g(p)}\frac{dHR}{dp} \qquad (9)$$

so that $g(p_k)$ can be recursively computed from $HR(p_k)$ and $FAR(p_k)$, $p_k=k/N$ (k=0 to N):

$$g(p_k) = (1-o)FAR(p_k) + oHR(p_k) - \sum_{k'>k} g(p_{k'}) \qquad (10)$$

$o(p_k)$ can then be recursively computed from $HR(p_k)$, $FAR(p_k)$ and $g(p_k)$:

$$o(p_k) = \frac{1}{g(p_k)}[oHR(p_k) - \sum_{k'>k} o(p_{k'})g(p_{k'})] \qquad (11)$$

Figure 6 shows the reliability curve obtained from an estimation of $o(p)$ and $g(p)$ through this method, for an anomaly above 1 (panel a) and 2 standard deviations (panel b) compared to that computed from the N+1 basic probability categories. In both cases the estimated curve looks very much like the curve one would have drawn manually for fitting the original curve.

Since the decision variable allowing to apply the bi-normal model might be close to log(p), as discussed at the end of the previous sub-section, the question arises of the suitability of directly fitting the reliability curve by applying a power relationship of the type $o(p)=\alpha p^{\beta}$. This model fits rather well the data and gives a reliability curve that is close to that obtained through the method described above, except for very rare events (not shown). It seems that a power model would have been a good choice for fitting the reliability curve, with respect to the data that have been considered in this study. However, this model might not be suitable for another set of data, while the method based on fitting the ROC curve can be applied in any case, provided the bi-normal model is applicable (which seems to be generally the case, as discussed in the previous sub-section).

(d) Quality of the estimation

For an anomaly above 1 standard deviation, lower probabilities are frequently forecast so that the lower part of the reliability curve shown in Fig. 1 (panel b) can be considered as reflecting the true performance of the forecast, at least up to p=6/51. The efficiency of the method of estimation described above has been investigated by comparing the lower part of the curve shown in Fig. 1 (panel b) to that obtained from a small sub-sample (M=1728) randomly extracted from the original data set (M=16848). The latter curve is very noisy, due to the small number of cases that have been considered in every probability category (except for

p=0). After applying the method of estimation the reliability curve is very close to that computed directly from the whole sample (Fig. 7).

Figure 8 (panel a) shows the reliability curve for an anomaly above 2.2 standard deviations. The original curve is very noisy, due to the rarity of the event (1.3% of the sample). The ROC curve fitting is good (correlation above 0.99), almost as good as that obtained for an anomaly above 1 standard deviation, although the sample is much smaller (not shown). But the credibility of the estimation of the reliability is questionable, in particular the fact that the right part of the curve tends to deteriorate, higher probabilities being more and more overestimated when the anomaly threshold increases, i.e. when the frequency of the considered event decreases (Fig. 6 and Fig. 7). This feature has been reported in earlier studies, e.g. by Buizza et al. (1999) for different thresholds of 12-hour precipitation amounts, and is also visible in Fig. 3.

For an anomaly above 2 standard deviations, Fig. 8 (panel b) shows the reliability curve that has been obtained by randomly halving the initial sample before applying the method of estimation. A comparison with Fig. 6 (panel b) shows that the estimation is close to that obtained from the whole sample. This does not prove that this estimation indicates the true performance of the forecast, but at least it shows that the method gives consistent results when the size of the sample is reduced to a certain extent. However, reducing the sample even further leads to increasing inconsistencies. Although the bi-normal model fits well the data even when the size of the sample is much reduced, as mentioned above, the right part of the curve tends to exhibit more and more the feature described in the previous paragraph, i.e. an increasing overestimation of higher probabilities. This last result suggests that this feature might be an artefact due to the limited size of the considered sample, rather than an intrinsic characteristic of the considered probabilistic forecast.

Figure 9 shows the reliability term of the Brier score when reducing gradually the size of the sample from 351 days to 16 days. The numerical value obtained through the method of estimation is not perfectly stable (solid line), but there is no tendency to underestimate the performance when the sample size decreases, contrary to what is the case when computing the reliability term from all 52 categories (dot-dashed line). On the other hand, the same stability can be achieved just by considering 10 probability categories instead of 52, except in the case of very small samples (dashed line). This indicates that the main advantage of the method lies in the possibility of estimating the reliability of probabilities that are not frequently forecast,

e.g. higher probabilities of rare events, rather than in a better estimation of the reliability term of the Brier score.

Note that the reliability term of the Brier score is weighted by the number of cases in each of the different probability categories that have been used for its estimation, so that in the case of rare events lower probabilities have a much larger impact than higher probabilities. One might consider, on the other hand, that the reliability of higher probabilities is an essential attribute of the quality of a probabilistic forecast, in general and especially for a rare event, just because it is in the case of higher probability that people really ask themselves: "Is this forecast reliable?". Lower probabilities of rare events may be disregarded most of the time by the general public, although these probabilities have a high value for certain users (Atger 2001). Therefore, the numerical value of the reliability term of the Brier score gives limited knowledge of the quality of probabilistic forecasts, as subjectively appreciated by end users. A careful examination of the full reliability curve is needed when one wants to get a meaningful picture of the performance. The method described in this section might help estimating reliability for infrequently forecast probabilities, even though there is obviously no guarantee that this estimation is valid when the considered sample is definitely too small.

6. Consequences for calibration

Several methods have been proposed for the calibration of ensemble forecasts (e.g. Hamill and Colucci 1998). The method used in the present study is based on reliability diagram statistics (e.g. Zhu et al. 1996). When i ensemble members forecast the event in an evaluation sample (i=0 to N), the calibrated probability is the frequency with which the event is observed when i ensemble members forecast the event in a calibration sample. This method has been applied to the data described in section 2, the calibration sub-sample and the evaluation sub-sample being obtained by halving randomly the available sample. Note that this scheme gives optimal results since both sub-samples are extracted from an identical population, which is not practically the case when calibrating probabilistic forecasts, due to the variability of the performance in time (Atger 2003).

Figure 10 shows the reliability diagram before (panel a) and after (panel b) calibration has been applied, for an anomaly above 1 standard deviation. The effect of fitting the data is also shown. Although the effect of calibration is rather noisy, the fitted curve indicates that reliability has been improved through calibration, even for higher, less sampled probability categories. This is not the case for an anomaly above 2 standard deviations (Fig. 11). Here the

calibration procedure is far less efficient: due to the limitation of the size of the sample, calibrating higher probabilities leads to a degradation of the reliability.

When the fitting procedure described in section 5 is applied to the calibration sub-sample, it is possible to define the calibrated probability as the *expected* frequency with which the event is observed when *i* ensemble members forecast the event. In other words, calibration is applied to fitted data. When this procedure is applied for an anomaly above 2 standard deviations the effect of calibration is positive for the main range of forecast probabilities, including higher probabilities (Fig. 12). This procedure also improves significantly the positive impact of calibration for an anomaly above 1 standard deviation, for the whole range of forecast probabilities (Fig. 13).

7. Conclusions

Different methods have been examined for the estimation of the reliability of ensemble-based probabilistic forecasts. ECMWF ensemble forecasts of the 850-hPa temperature anomaly over Europe have been accumulated for 4 winter seasons (351 days) in order to get a sufficiently large verification sample.

The estimation of reliability is traditionally achieved by considering a certain number of probability categories. Due to the relatively small size of available samples, the reliability term of the Brier score increases with the number of probability categories that are considered for its computation. The overall performance, in terms of reliability, is thus likely to be wrongly estimated. Furthermore, the reliability of higher probabilities, in the case of forecasting rare events, cannot be estimated from the considered sample.

A significance testing procedure has been designed in order to determine whether consecutive probability categories should be merged into a single one for estimating reliability. For events occurring frequently (e.g. an anomaly above 1 standard deviation) this method leads to credible estimates of the reliability for the whole range of forecast probabilities. On the other hand, in the case of rare events, this method does not allow an estimation of the reliability of higher probabilities.

A parameterisation scheme has been designed in order to estimate reliability from limited samples, even in the case of rare events. The procedure consists in fitting an ROC curve under the bi-normal assumption. The quality of the fitting appears very high. Once the ROC curve is fitted, reliability is estimated through a recursive computation of the expected frequencies

with which (i) probabilities would be forecast, and (ii) the event would be observed, in an infinite, hypothetical sample. This estimation leads to a very good fit of the original reliability curve.

The validity of the method has been demonstrated, to a certain extent, by estimating the reliability from a small sub-sample randomly extracted from the original set of data. For lower probabilities, that are frequently forecast, this estimation is close to the true performance as evaluated directly from the whole sample. However, the estimation of the reliability of probabilities that have been forecast in only a few occasions remains questionable, e.g. when considering very rare events and/or higher probabilities.

Consequences for the calibration of probabilistic forecasts have been investigated, in particular in the case of rare events. Using a basic method of calibration, it has been shown that the positive impact on reliability is increased, especially for higher probabilities, when the calibrated probability is estimated after fitting the ROC curve, rather than computed directly from the observed frequency of occurrence.

The reasons why the bi-normal model allows an almost perfect fit of ensemble-based ROC curves remain unclear. About the apparently universal appropriateness of the bi-normal model in fitting ROC curves, Hanley (1988) reviews a number of justifications that have been proposed in previous studies. He concludes that the robustness of the bi-normal assumption is not an intrinsic characteristic of the ROC curve, but might be a mere consequence of (i) sampling limitations that prevent the lack of fit to be significantly attributed to the bi-normal model, and (ii) the limited number of rating categories, i.e. the number of points forming the ROC curve. None of these two arguments is valid in the present study, given the quality of the fit and the large number of probability categories that have been considered for constructing the ROC curves.

The most convincing explanation invokes the central limit theorem (e.g. Green and Swets 1966). Assuming the considered event is the effect of a large number of independent causes, the underlying decision variable may be the sum of a large number of independent random variables. In this case, the signal and noise distributions are Gaussian just because they are sums of conditional distributions of independent variables. This hypothesis is supported by preliminary results (not shown) suggesting that the binormal model fails, in certain circumstances, to fit data derived from idealized experiments in which the occurrence of the considered event is entirely described by a unique random variable. Further work is however

needed for a confirmation of this result.

Acknowledgement

(…)

References

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 9, 1941-1953.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes Geophys.*, **8**, 401-417.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts - Consequences for calibration. *Mon. Wea. Rev.*, accepted.

Déqué, M., 2001: Seasonal predictability of tropical rainfall: probabilistic formulation and validation. *Tellus*, **53A**, 500-512.

Green, D. M. and J. A. Swets, 1966: Signal detection theory and psychophysics. New York, Wiley and Sons.

Hamill, T. M, 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

Hamill, T. M., and S. J. Colucci, 1998: Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Hanley, J. A., 1988: The robustness of the binormal assumptions used in fitting ROC curves. *Med. Decis. Making*, **8**, 197-203.

Harvey, L.O., K.R. Hammond, C.M. Lusk & E.F. Mross, 1992. The application of signal detection theory to weather forecasting behaviour. *Mon. Wea. Rev.*, **120**, 863-883.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for Ensemble Prediction Systems. *Wea. Forecasting*, **15**, 5, 559-570.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie and H.L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Kobayashi, C., K. Yoshimatsu, S. Maeda, and K. Takano, 1996: Dynamical one-month forecasting at JMA. Preprints, 11th Conf. on Numerical Weather Prediction, Norfolk, VA, Amer. Meteor. Soc., 13-14.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Mullen, S. L. and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-661.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595-600.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

Murphy, A. H. and R.L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tribbia, 1993: Ensemble prediction. Seminar on Validation of Models over Europe, Reading, U.K., European Centre for Medium-range Weather Forecasts (Proceedings, vol. 1, 21-66).

Richardson, D. S, 2000: Skill and economic value of the ECMWF Ensemble Prediction System, *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.

Swets, J. A., 1986: Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psycholog. Bull.*, **99**, 181-198.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts. Proceedings, 1-25.

Toth Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Tracton M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting*, **8**, 379-398.

Zhu, Y., G. Yyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conference on Weather Analysis and Forecasting, Norfolk, Virginia, Amer. Meteor. Soc., J79-J82.

Figures

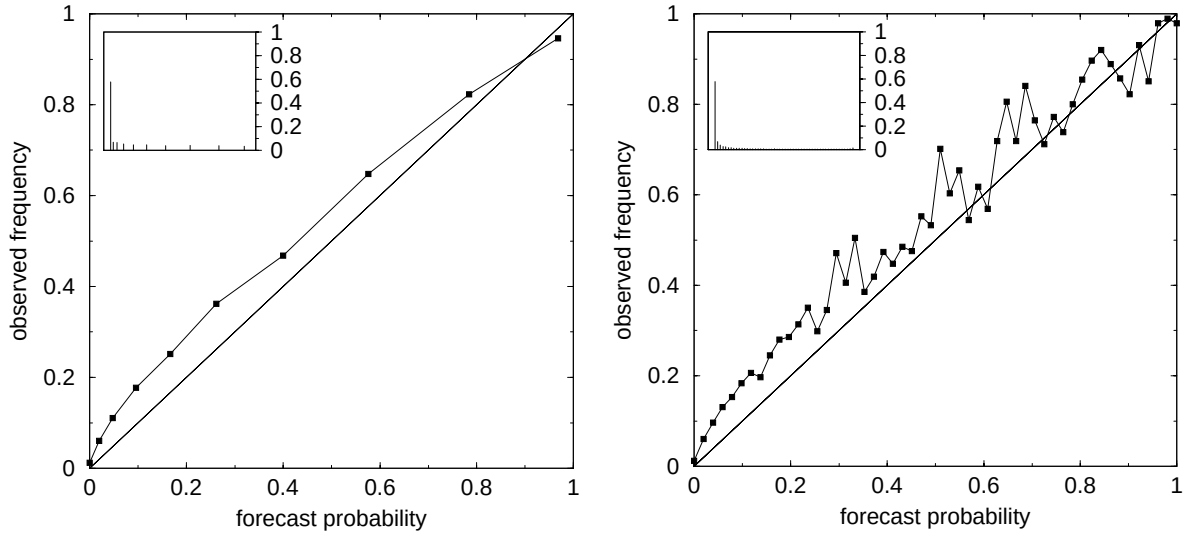


Figure 1. (main graph) Reliability curve for an anomaly above 1 standard deviation. (inset) Proportion of cases in each probability category. (a) 10 probability categories. (b) 52 probability categories, i.e. all probability categories that can be defined from a 51-member ensemble.

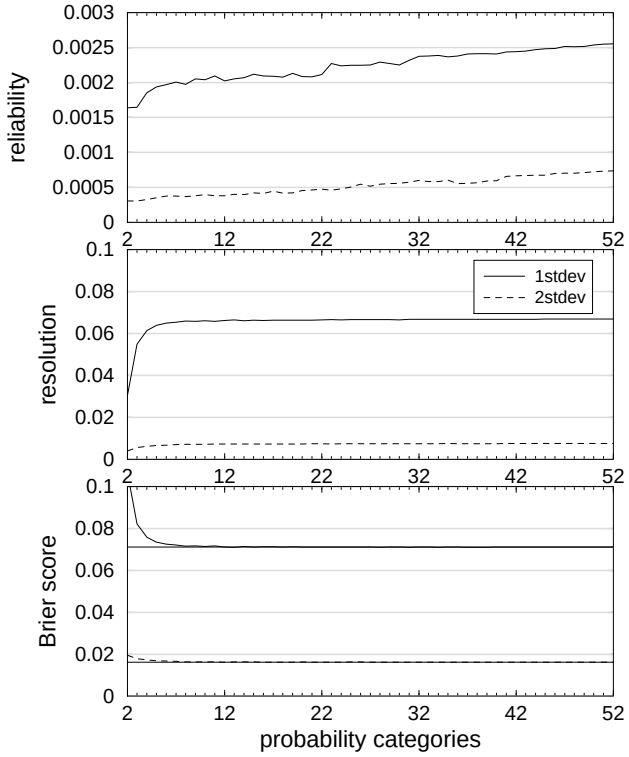
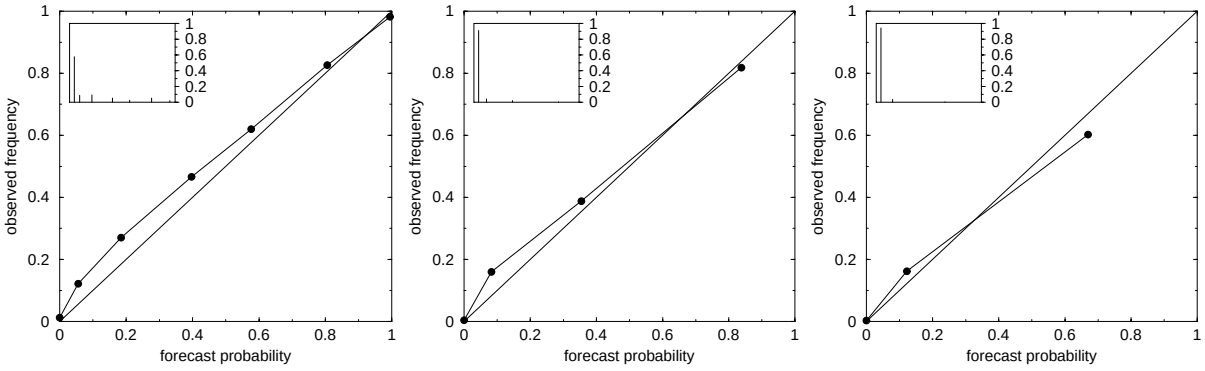


Figure 2. (a) (top) Reliability term of the Brier score for an anomaly above 1 standard deviation (solid line) and 2 standard deviations (dashed line) as a function of the number of probability categories. (b) (middle) Same as (a) but for the resolution term of the Brier score. (c) (bottom) Same as (a) but for the Brier score computed from the reliability and resolution terms after decomposition. Horizontal thin solid lines indicate the true value of the Brier score.



(a) (b) (c)

Figure 3. (a) Same as Fig. 1, but for a number of probability categories that has been defined through a significance testing procedure (see text). (b) Same as (a) but for an anomaly above 2 standard deviations. (c) Same as (a) but for an anomaly above 2.2 standard deviations.

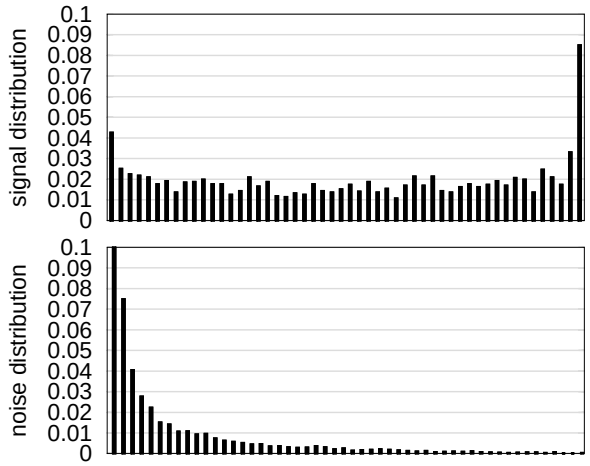


Figure 4.

(a) (top) Signal distribution, i.e. the distribution of the forecast probability i/N, $i=0$ to N, when the anomaly is above 1 standard deviation.

(b) (bottom) Same as (a) but for the noise distribution, i.e. the distribution of the forecast probability i/N, $i=0$ to N, when the anomaly is not above 1 standard deviation (first bar peaks at 0.6).

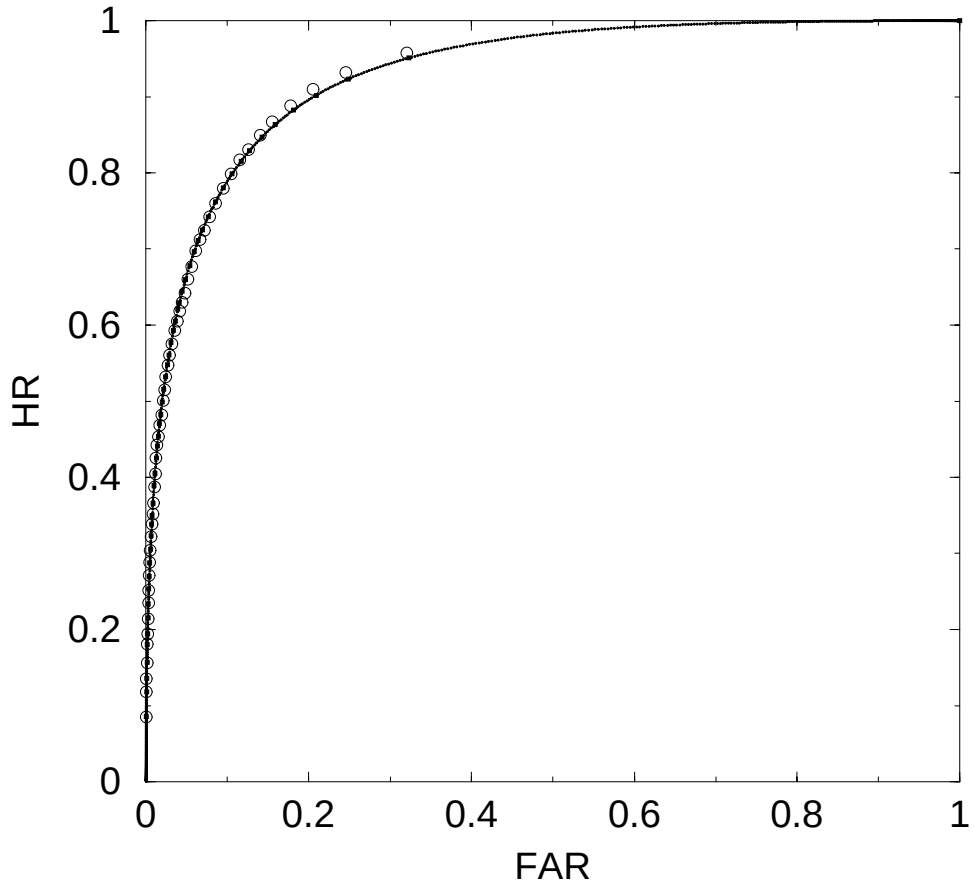


Figure 5 (a) ROC curve for an anomaly above 1 standard deviation (large circles), ROC curve obtained after fitting the bi-normal model to the data and sampling with 1000 points (thin dots), orthogonal projection of the original ROC points onto the fitted curve (thick dots).

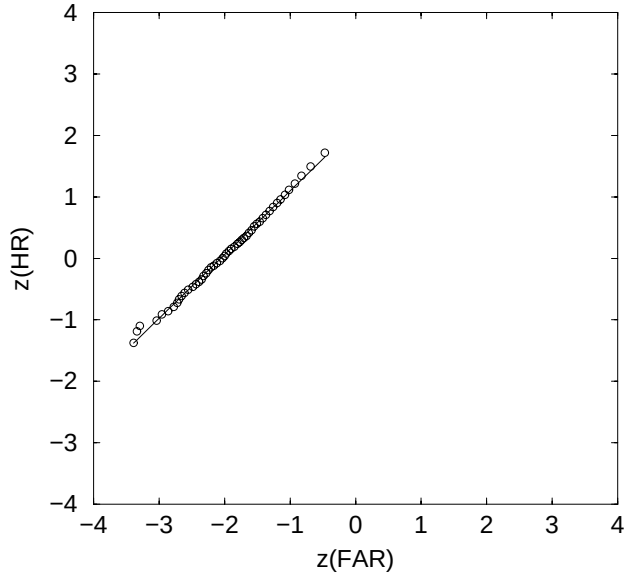


Figure 5. (b) ROC curve for an anomaly above 1 standard deviation after transformation of *HR* and *FAR* into their corresponding standardized normal deviates (circles). Linear fit, correlation 0.99 (solid line).

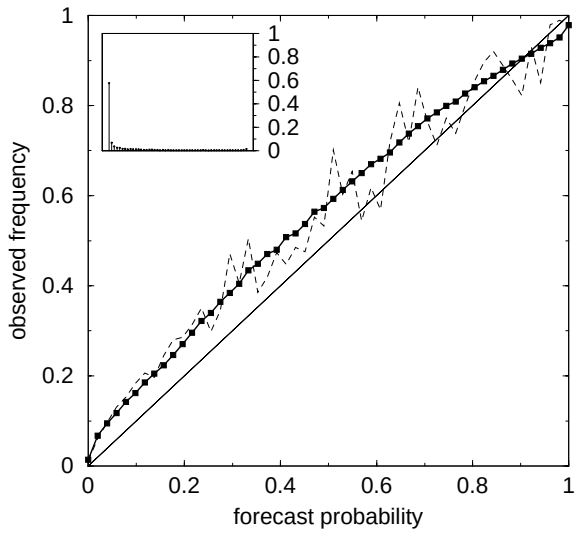


Figure 6 (a)

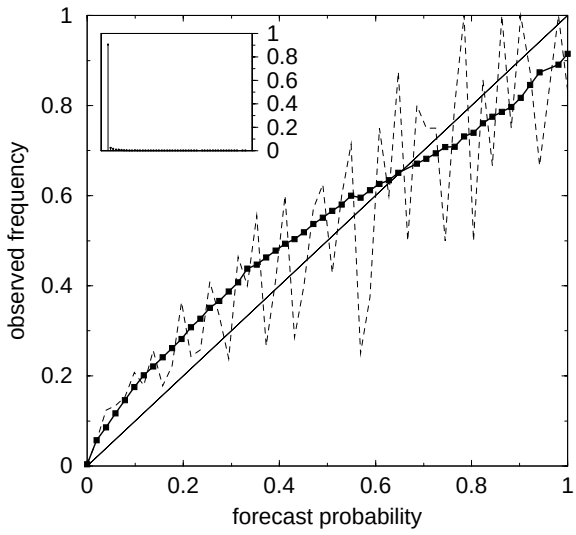


Figure 6 (b)

Figure 6. Reliability curve computed from the basic 52 probability categories (dashed line) and as obtained after fitting the ROC curve and estimating *o(p)* and *g(p)* (solid line and squares). (a): Anomaly above 1 standard deviation. (b): Anomaly above 2 standard deviations.

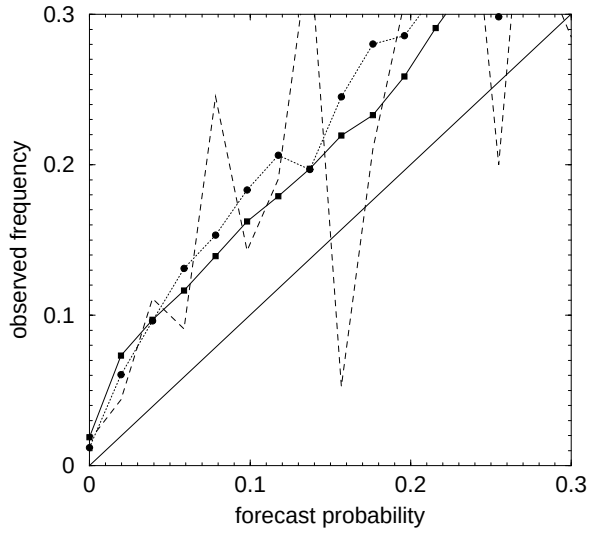
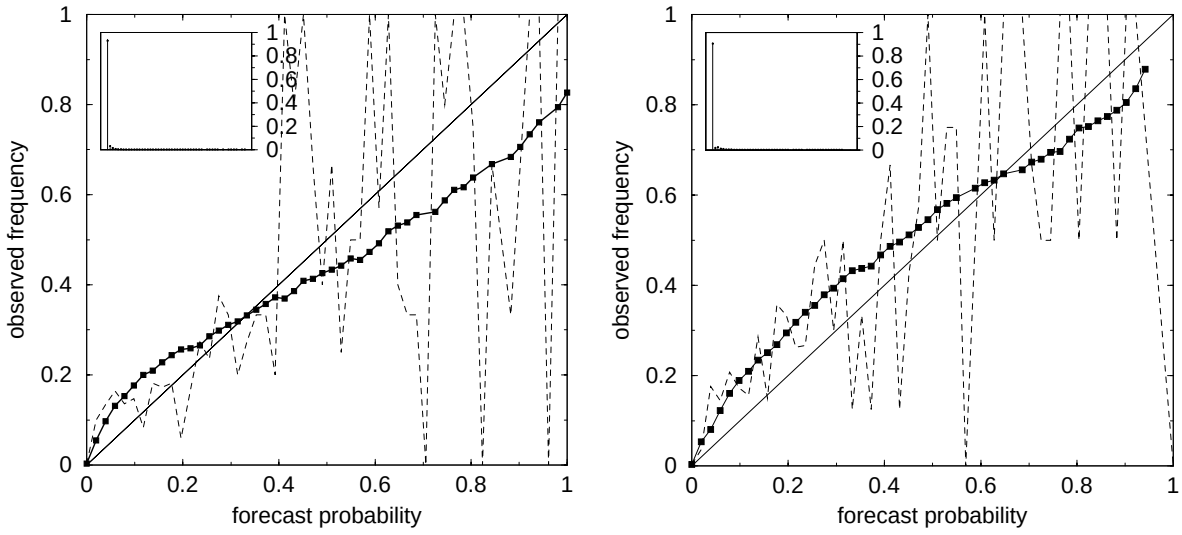


Figure 7. Reliability curve for an anomaly above 1 standard deviation. Computed from 52 probability categories, full sample (M=16848) (dotted line, circles), reduced sample (M=1728) (dashed line). Obtained after fitting the ROC curve, reduced sample (M=1728) (full line, squares). Only the lower part of the curve is shown ($p<0.3$)



(a) (b)

Figure 8. Same as Fig. 6, but (a): Anomaly above 2.2 standard deviations. (b): Anomaly above 2 standard deviations, the sample has been randomly halved.

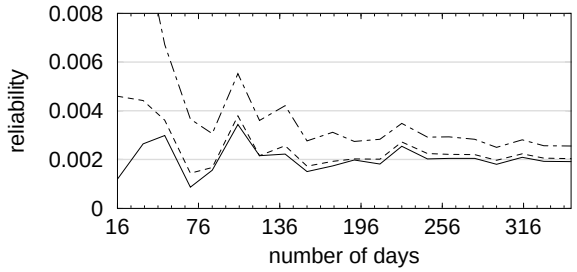
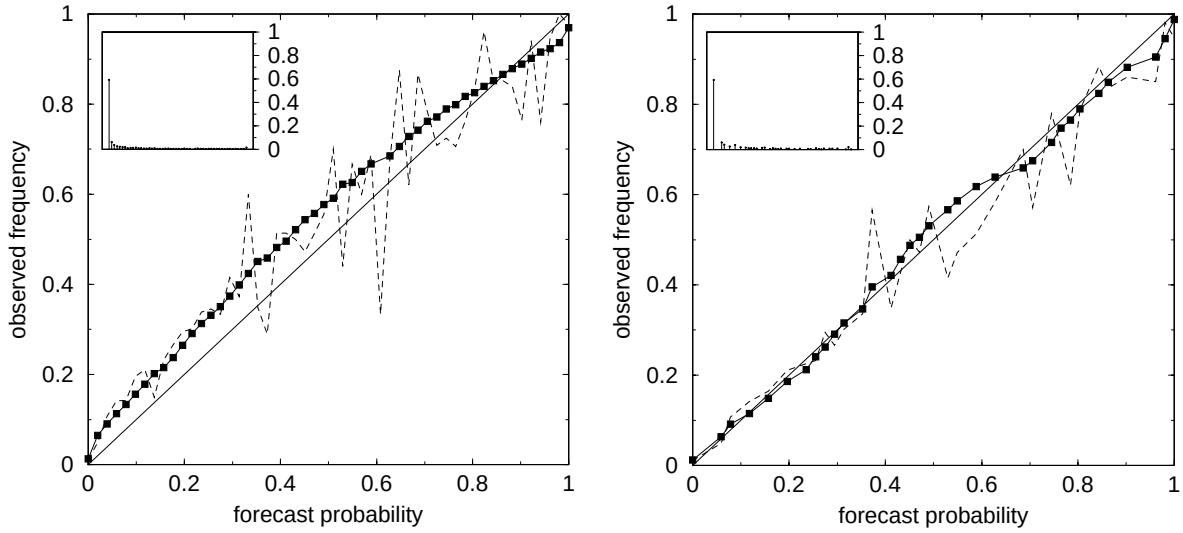
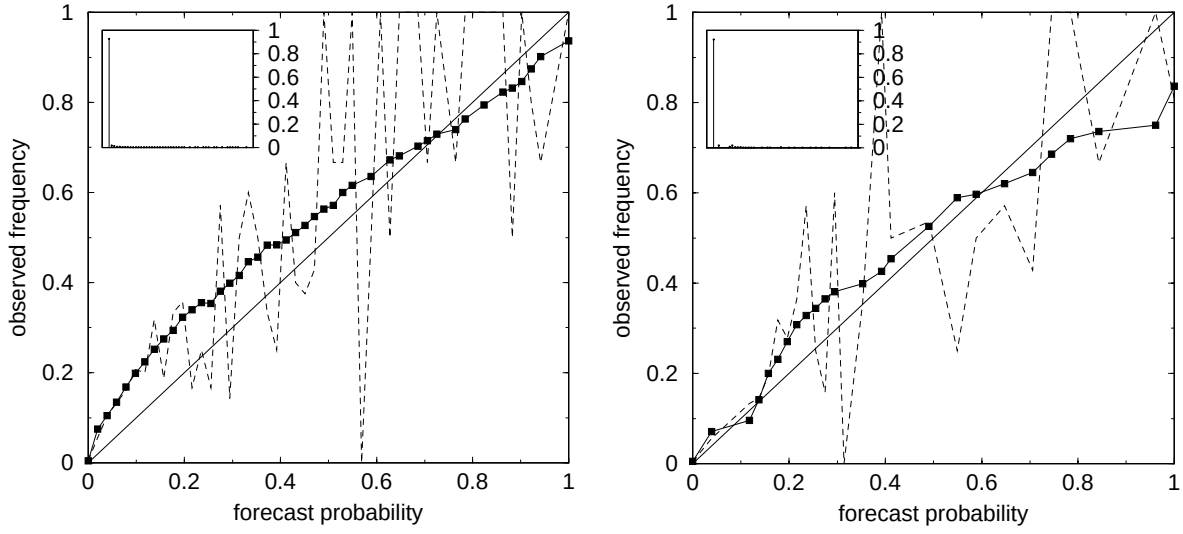


Figure 9. Reliability term of the Brier score for an anomaly above 1 standard deviation, as a function of the sample size. 52 probability categories (dot-dashed line), 10 probability categories (dashed line), after fitting an ROC curve (solid line).



(a) (b)

Figure 10. Reliability curve for an anomaly above 1 standard deviation, computed from 52 probability categories (dashed line) and as obtained after fitting the ROC curve and estimating *o(p)* and *g(p)* (solid line). The sample has been halved randomly. (a) Before calibration. (b) After calibration based on 52 probability categories.



(a) (b)

Figure 11. Same as Fig. 10 but for an anomaly above 2 standard deviations.

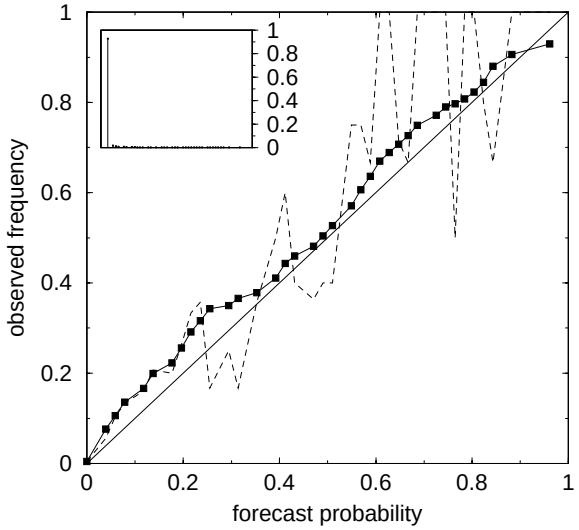


Figure 12. Same as Fig. 11(b) but calibration is based on fitted data (see text).

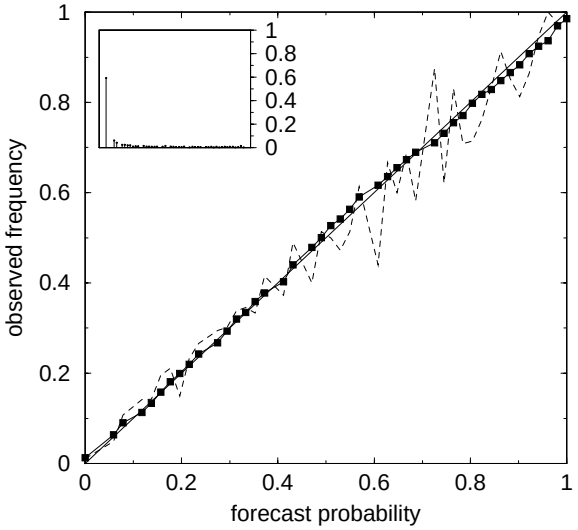


Figure 13. Same as Fig. 12 but for an anomaly above 1 standard deviation.

Conclusion et perspectives

Les différents travaux présentés dans cette thèse abordent tous, de manière plus ou moins directe, la question suivante : « Qu'est ce qu'un bon système de prévision d'ensemble ? » La réponse à cette question se trouve déjà, paradoxalement, dans le premier chapitre de la thèse : un bon ensemble est fiable, et il possède en même temps une bonne dose de résolution. Fiabilité et résolution sont à prendre ici dans le sens précis qui leur est donné dans le chapitre 1 (pour une prévision probabiliste), puis dans le chapitre 3 (pour un ensemble). Il semble exister une forme d'exclusion entre ces deux attributs de la qualité d'un système de prévision d'ensemble. Un ensemble composé d'observations extraites au hasard d'une archive climatologique représentative de la période considérée est parfaitement fiable mais ne présente aucune résolution (Talagrand et al., 1997). Au contraire, un système de prévision déterministe peut avoir une résolution élevée, directement en rapport avec la performance du modèle utilisé, tandis que sa fiabilité reste médiocre (sauf dans le cas d'une prévision parfaite) du fait que seuls des Diracs sont prévus tandis que les distributions observées correspondantes présentent une dispersion d'autant plus élevée que la performance est limitée. Un bon ensemble est donc issu d'un compromis entre ces deux extrêmes : suffisamment probabiliste pour être fiable, suffisamment déterministe pour ne pas perdre toute sa résolution.

§§§

Les chapitres 1 et 2 tentent de répondre à une première question complémentaire : « Que doit-on attendre d'un système opérationnel de prévision d'ensemble ? » Les principaux critères examinés sont les suivants : (i) L'ensemble doit fournir des prévisions probabilistes de qualité supérieure à celles obtenues directement à partir de sa prévision de contrôle (sinon, ce n'est pas la peine de construire l'ensemble); (ii) L'ensemble doit être plus performant qu'un ensemble « du pauvre » obtenu à moindre coût (sinon, il faut suivre une autre stratégie

pour la construction de l'ensemble); (iii) La performance de l'ensemble doit décroître de façon significative quand on limite le nombre de ses membres (sinon, ce n'est pas la peine de disposer d'autant de membres); (iv) Les prévisions probabilistes issues de l'ensemble doivent posséder une valeur économique, au moins potentielle (sinon, l'utilité pratique de l'ensemble est discutable).

Les chapitres 3 et 4 abordent une deuxième question complémentaire, beaucoup plus ciblée que la précédente, mais qui a des répercussions pratiques importantes : « Comment mesurer de manière fiable la performance de prévisions probabilistes issues d'un système de prévision d'ensemble ? » Les deux problèmes abordés sont d'une part celui de la catégorisation des probabilités prévues, qui est à la base de la décomposition du score de Brier, d'autre part celui de la stratification des données, rendue nécessaire par la petitesse des échantillons disponibles pour la vérification. Dans les deux cas, les travaux présentés conduisent à des recommandations portant non seulement sur l'évaluation, mais aussi sur la calibration de prévisions probabilistes issues d'un ensemble.

§§§

Une question plus fondamentale que les deux précédentes apparaît en filigrane dans les 3 premiers chapitres de cette thèse : « Qu'est ce qui *fait* un bon système de prévision d'ensemble ? ». Un ensemble n'est pas un système quelconque de prévision probabiliste. Il est construit à partir d'un modèle, utilisé non seulement pour effectuer des prévisions à partir d'états initiaux perturbés, mais aussi au sein du système d'assimilation qui permet d'établir l'état initial de référence. Il va de soi que les caractéristiques du modèle sous-jacent ont une influence sur la qualité d'un ensemble. Par ailleurs, on ne peut douter que l'ensemble possède des caractéristiques propres, indépendantes du modèle de prévision, qui sont la conséquence des techniques utilisées pour générer ses membres.

La validation d'un système de prévision repose en partie sur la comparaison de sa performance avec celle de systèmes dans lesquels les choix de mise en œuvre

opérationnelle sont différents. Dans le cas d'un ensemble, la méthode de génération des perturbations initiales, par exemple, est supposée être à l'origine de certaines différences de performance. La qualité du modèle sous-jacent est cependant un facteur d'influence majeur. Il serait par conséquent périlleux de conclure à la supériorité d'une approche ensembliste par rapport à une autre, dès lors que les modèles sous-jacents présentent des différences de performance. Ce type de comparaison est cependant fréquent (Mullen & Buizza, 2001). Les résultats d'une analyse de l'impact des caractéristiques respectives de l'ensemble et du modèle sous-jacent, sur la performance de prévisions probabilistes, seraient de nature à éclairer les résultats obtenus dans ces études.

§§§

La définition la plus générale qu'on puisse donner de la fiabilité d'un système de prévision d'ensemble est la suivante : c'est la correspondance entre une loi de distribution ϕ et la distribution de la vérification dans les cas où l'ensemble est distribué suivant ϕ. La loi de distribution ϕ peut être caractérisée par ses différents moments statistiques, et la fiabilité peut ainsi être définie comme une correspondance entre ces moments. Si on néglige les moments supérieurs la fiabilité se résume ainsi à la correspondance des moyennes et des variances. Evaluer la fiabilité revient donc à estimer des biais (conditionnels) de positionnement et de dispersion de l'ensemble. Les membres d'un ensemble étant des intégrations d'un même modèle, une hypothèse simplificatrice[9] est qu'ils présentent les mêmes biais. Par conséquent les biais de positionnement de l'ensemble sont les biais du modèle sous-jacent, et ne dépendent pas de la méthode de génération des membres de l'ensemble. Quant aux biais de dispersion, ils reflètent au contraire une déficience propre au système de prévision d'ensemble et indépendante du modèle.

[9] On néglige ici les effets d'éventuelles perturbations de la formulation du modèle utilisé pour générer les membres de l'ensemble. On néglige également les biais conditionnels dont l'origine réside dans les imperfections de la procédure de perturbation de l'état initial.

La résolution est définie comme la variabilité de la distribution de la vérification quand la distribution de l'ensemble varie. Si on néglige à nouveau les moments d'ordre supérieur, et qu'on se place dans le cadre d'une fiabilité parfaite, la résolution indique la variabilité de la moyenne et de la dispersion de l'ensemble. Le second terme reflète une caractéristique de l'ensemble. Quant au premier terme, il peut être considéré comme une mesure du niveau moyen de performance de la moyenne de l'ensemble (en l'absence de biais conditionnels, seule une prévision parfaite varie autant que la vérification), dont on peut penser qu'elle dépend largement de la performance du modèle sous-jacent.

§§§

Dans des expériences récentes, nous avons tenté de générer des ensembles suivant un modèle statistique permettant de reproduire les principales caractéristiques des ensembles opérationnels. En faisant varier les paramètres de la simulation on peut évaluer l'impact, sur la performance de prévisions probabilistes, des différents facteurs énoncés ci-dessus, certains attribuables au seul modèle de prévision, d'autres représentant des caractéristiques propres à l'ensemble.

Les premiers résultats semblent indiquer que le terme de fiabilité du score de Brier dépend d'abord des biais du modèle sous-jacent, et dans une moindre mesure seulement des biais de dispersion de l'ensemble. Ce résultat permet d'éclairer une des conclusions du chapitre 3 qui est que la variabilité spatiale et temporelle du biais systématique du modèle contribue pour une large part aux variations de fiabilité de l'ensemble.

En ce qui concerne le terme de résolution, il semble largement conditionné par le niveau moyen de performance de la moyenne d'ensemble, la variabilité de la dispersion n'y contribuant que très faiblement. Un résultat du chapitre 1 est que la résolution de l'ensemble du CEPMMT provient essentiellement de sa moyenne, tandis que sa dispersion n'y contribue que marginalement. Ce résultat a pu être

interprété comme le signe d'une qualité médiocre de la dispersion. Cette interprétation pourrait être remise en question, puisqu'il semble que ce soit par construction que la dispersion a peu d'effet sur le terme de résolution du score de Brier (au moins pour des systèmes dont les caractères quantitatifs sont ceux des systèmes actuels de prévision météorologique à grande échelle). Il s'agit d'ailleurs d'un résultat assez intuitif : le premier moment de la distribution a un poids beaucoup plus important que le second moment. Par conséquent, le terme de résolution du score de Brier[10] pourrait ne pas être un critère de choix pour apprécier les effets d'une amélioration de la méthode utilisée pour générer un ensemble.

On mentionne cependant dans le chapitre 1 que la dispersion (au sens qualitatif cette fois) contribue indirectement à la résolution de l'ensemble du CEPMMT puisque c'est elle qui permet d'avoir une moyenne d'ensemble plus proche de la vérification, en moyenne, que la prévision de contrôle. Par définition, l'EQM de la moyenne d'ensemble est égale à l'EQM des membres de l'ensemble, diminuée de la variance de l'ensemble. Pour générer un ensemble à forte résolution à partir d'un état initial de référence, il faut donc trouver un compromis entre une forte dispersion, qui éloigne les membres de la prévision de contrôle et dégrade donc leur performance moyenne (puisque la prévision de contrôle est obtenue à partir d'une estimation supposée optimale des conditions initiales), et une faible dispersion qui rapproche les membres de la prévision de contrôle mais dégrade la performance de la moyenne d'ensemble.

Dans le cas d'un ensemble « du pauvre », examiné dans les chapitres 1 et 2, c'est toujours la qualité de la moyenne d'ensemble qui conditionne la résolution (et donc aussi la valeur économique potentielle, qui est une mesure de la résolution). La variance moyenne d'un tel ensemble est généralement faible, en raison du petit nombre de prévisions qui peuvent être réunies. Si cependant sa résolution

[10] Et donc aussi le score de Brier, qui est généralement dominé par le terme de résolution (chapitre 4).

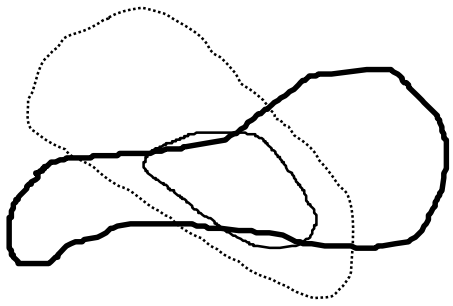
dépasse celle d'un ensemble opérationnel (comme on le constate en pratique), c'est que le niveau moyen de performance des membres est plus élevé. Ceci n'a rien d'étonnant, puisque chacun des membres constituant un tel ensemble est une prévision intégrée à partir d'un état initial supposé optimal pour le modèle considéré, et non à partir d'un état initial statistiquement dégradé par rapport à l'état initial de référence.

Une des faiblesses des ensembles opérationnels pourrait ainsi être la forte proportion de membres qui s'éloignent de la vérification, du fait d'une trop grande amplitude des perturbations initiales. Le problème majeur des ensembles n'est peut-être pas qu'ils sont sous-dispersifs, mais plutôt qu'ils échouent à échantillonner la pdf comme il le faudrait. Il s'agirait d'un problème de positionnement de l'ensemble, davantage que d'un problème de dispersion. D'un point de vue pratique, et gardant à l'esprit qu'un ensemble sert avant tout à indiquer à un prévisionniste les évolutions atmosphériques envisageables, il pourrait même être préférable de limiter la dispersion afin que la distribution échantillonne une partie seulement de la pdf, mais ne « déborde » pas au-delà. Connaissant le niveau de sous-dispersion on saurait alors que seule une fraction des possibilités est représentée par l'ensemble, mais au moins éviterait-on de considérer des alternatives improbables car n'étant pas représentatives de la pdf (voir schéma ci-dessous).

Trait gras : ensemble parfait (pdf)

Trait tireté : ensemble non sous-dispersif mais mal positionné

Trait fin : ensemble très sous-dispersif mais correctement positionné



§§§

Le schéma reproduit ci-dessus laisse penser qu'il existe, a priori, une pdf de l'état futur de l'atmosphère qu'un ensemble s'efforce d'échantillonner. Par conséquent,

étant donné un état initial, il existe une probabilité définie *a priori* qu'un événement météorologique donné se produise. On pourrait voir dans cette proposition un argument pour la théorie de la propension de Popper (1959). Opposé à l'interprétation subjectiviste de la probabilité, qui en fait une mesure de notre méconnaissance du réel, Popper propose de considérer la probabilité comme un attribut de l'objet auquel elle s'applique, et plus précisément comme une mesure de sa propension ('propensity') à se trouver dans un état particulier. La probabilité *a priori* qu'implique la discussion qui précède n'est cependant pas une propension, car elle n'est définie que dans le cadre strict d'un système de prévision, et reste relative à la qualité de ce système. C'est l'incertitude de prévision du modèle que l'ensemble cherche à évaluer, et non une incertitude qui serait intrinsèque à l'atmosphère. Cette probabilité « objective » est tout à fait compatible avec l'interprétation subjectiviste, la probabilité subjective n'exprimant rien d'autre que le niveau d'incertitude d'une prévision élaborée par le « système de prévision » très particulier qu'est le prévisionniste.

§§§

Plus haut dans cette conclusion il est fait mention du fait que fiabilité et résolution peuvent s'exprimer sous la forme de relations entre les moments statistiques des distributions prévue et observée. Cette approche permet d'aborder de manière originale l'évaluation de la performance d'un système de prévision d'ensemble, sans passer par la vérification de prévisions probabilistes. La courbe de fiabilité trouve son équivalent dans une série de diagrammes mettant en évidence la correspondance entre les différents moments. La fiabilité peut être mesurée par l'écart entre distributions prévue et observée, après classification en fonction des différents moments. La résolution peut être mesurée par la variabilité de la « distribution prévisible », suivant une telle classification. En pratique, on peut se contenter d'évaluer la performance relative aux deux premiers moments, et vérifier ainsi que la performance de la moyenne conditionne largement la qualité d'un ensemble. Cette approche permet

également de considérer d'une manière nouvelle la 'spread-skill relationship', propriété qui possède ici une définition précise, celle de la correspondance entre les moments d'ordre 2. L'examen des moments supérieurs permet enfin d'étudier la pertinence des « détails » qu'un ensemble opérationnel formé de plusieurs dizaines de membres peut apporter à une distribution, en particulier la dissymétrie et la multimodalité, et plus généralement tout caractère non Gaussien (Denholm-Price, 2003).

Bibliographie

Denholm-Price, J.C.W., 2003. Can an ensemble give anything more than Gaussian probabilities? *Nonlinear Processes in Geophysics*, submitted.

Mullen, S. L. and R. Buizza, 2001. Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-661.

Popper, K., 1959 : The propensity interpretation of probability. *British Journal of the Philosophy of Science*, **10**, 25-42.

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Seminar on Predictability, Reading, U.K., European Centre for Medium-range Weather Forecasts. Proceedings, 1-25.

Validation and some properties of meteorological ensemble prediction systems

Probabilistic meteorological forecasts based on ensemble prediction systems are evaluated. The resolution and reliability components of the decomposition of the Brier score are used for quantifying the performance. Probabilistic forecasts based on operational ensembles are compared to those obtained from a single model run, through a statistical scheme. « Poorman ensembles », consisting of a few deterministic forecasts run in different operational centres, are evaluated too. The conditions for a realistic estimation of the performance of ensemble based probabilistic forecasts are also investigated. The spatial and interannual variability of the reliability implies a strong stratification of the data, that is not always possible with available samples limited in size. Another, essential issue is the categorization of forecast probabilities, required for achieving the decomposition of the Brier score.

AUTEUR : Frédéric Atger

TITRE : Validation et étude de quelques propriétés de systèmes de prévision météorologique ensemblistes

DIRECTEUR DE THESE : Olivier Talagrand

LIEU ET DATE DE SOUTENANCE : 14 avril 2003 à Toulouse

RESUME :

Les travaux présentés portent sur l'évaluation de prévisions météorologiques probabilistes issues de systèmes de prévision d'ensemble. Les principaux critères utilisés sont les termes de résolution et de fiabilité du score de Brier. Les prévisions issues de systèmes opérationnels sont comparées à celles obtenues par des méthodes statistiques à partir de l'intégration unique d'un modèle de prévision. On s'intéresse également à des systèmes de prévision d'ensemble consistant à regrouper les prévisions issues de quelques centres opérationnels. Les conditions requises pour une estimation réaliste de la performance de prévisions issues d'un ensemble sont examinées par ailleurs. La variabilité spatiale et temporelle de la fiabilité impose une stratification des données que ne permettent pas toujours les échantillons de taille réduite disponibles pour la vérification. Un autre problème essentiel est celui de la catégorisation des probabilités prévues, qui permet la décomposition du score de Brier.

MOTS-CLES : prévision d'ensemble – prévision probabiliste – qualité des prévisions – score de Brier

DISCIPLINE ADMINISTRATIVE : météorologie

INTITULE ET ADRESSE DU LABORATOIRE :
Centre National de Recherches Météorologiques – Météo-France – 42, avenue G. Coriolis – 31057 Toulouse cedex