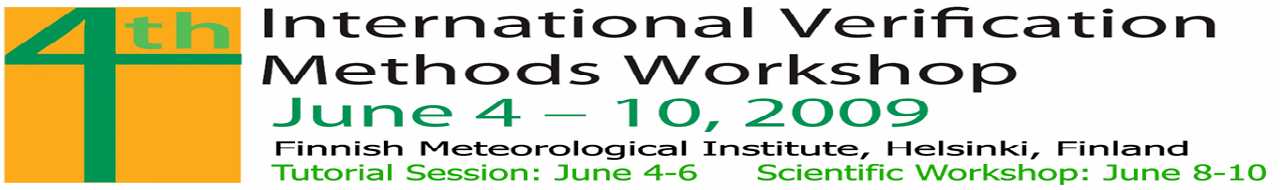

di Arturo Pucillo
OSMER - ARPA FVG

"4th international methods tutorial" FMI, Helsinki, 4-6 June 2009

Lo scopo dichiarato del workshop è la valutazione dello stato dell'arte nelle tecniche più avanzate di verifica delle previsioni e la condivisione di lavori di ricerca e di applicazione attualmente in uso nella comunità scientifica. Il team organizzatore é costituito da un bel gruppo di persone che si occupano di verifica delle previsioni dagli "albori", e attualmente rappresenta la "testa di ponte" verso le nuove tecniche via via inventate, teorizzate e sviluppate: **Barbara Brown** dell'NCAR, **Elizabeth Ebert** dell'Australian Meteorology Bureau, **Ian Jolliffe** dell'università di Exeter, **Martin Goebber** del DWD, **Barbara Casati** di Ouranos (Canada), **Laurie Wilson** del Meteorological Service of Canada, **Anna Ghelli** dell'ECMWF, **Pertti Nurmi** dell'FMI (local organizer), **Matthew Pocernich** dell'NCAR. L'impressione che se ne ricava, suffragata da colloqui avuti nel corso della settimana con queste persone, é di un gruppo molto affiatato in un campo che ancora non è del tutto maturo dal punto di vista scientifico. Da qui é nato il desiderio di impegnarsi, anche al di là dei doveri professionali di ciascun membro del "lead group", nella divulgazione e nella didattica, anche attraverso la forma del "tutorial" come quello a cui ho partecipato dal 4 al 6 giugno, nei giorni precedenti il workshop.
Questo tutorial, a cui hanno avuto accesso 26 persone da 25 Paesi diversi, si é articolato in alcune lezioni che hanno trattato gli argomenti fondamentali della forecast verification, unitamente ad alcune sessioni dedicate allo sviluppo di un progetto di verifica su opportuni dataset di osservazioni e simulazioni relativi a problematiche diverse, il tutto per acquisire confidenza nell'individuazione degli obiettivi principali di verifica e delle relative tecniche, utilizzando gli strumenti "free", sviluppati dai tutor del corso, nell'ambito del software statistico open source R.

Che vantaggi ne ho tratto personalmente (e di conseguenza ne ha tratto l'OSMER)?
Una chiarificazione generale sulla struttura dei problemi di verifica delle previsioni (obiettivi e tecniche associate), una conoscenza più approfondita di alcune tecniche interessanti per il mio lavoro all'OSMER, una maggior dimestichezza con R e con i "packages" di verifica, la conoscenza personale (e buone relazioni) con chi si occupa di verifica delle previsioni in giro per il mondo, la possibilità di contatti futuri con chi affronta problemi di verifica simili ai miei.

La seconda parte della missione ha visto la mia partecipazione al workshop, dove ho seguito le presentazioni di diverse persone provenienti da ogni parte del mondo su 9 sessioni più una dedicata alle presentazioni dei progetti di verifica sviluppati durante la Tutorial Session. A questa sessione ho partecipato attivamente in quanto incaricato dal mio gruppo di lavoro di presentare il nostro progetto di verifica.

Tutorial Session.

Barbara Brown: Basic Verification Concepts

Il concetto di verifica (sinonimo di conferma) delle previsioni si vede applicato a diversi ambiti: amministrativo, scientifico, economico. La verifica diventa uno degli strumenti per stabilire la "bontà" (goodness) di una previsione, e in generale ne misura la qualità (che è cosa diversa dal valore -ad esempio economico-: paradossalmente in certi casi valore e qualità sono inversamente proporzionali).

Per affrontare correttamente un problema di verifica, è necessario porsi degli obiettivi, ovvero delle domande, del tipo: in che zone il modello ha la migliore performance? Ci sono dei regimi di circolazione atmosferica in cui le previsioni sono migliori o peggiori? La previsione probabilistica è ben calibrata (ossia affidabile)? Le previsioni rappresentano bene la variabilità meteorologica

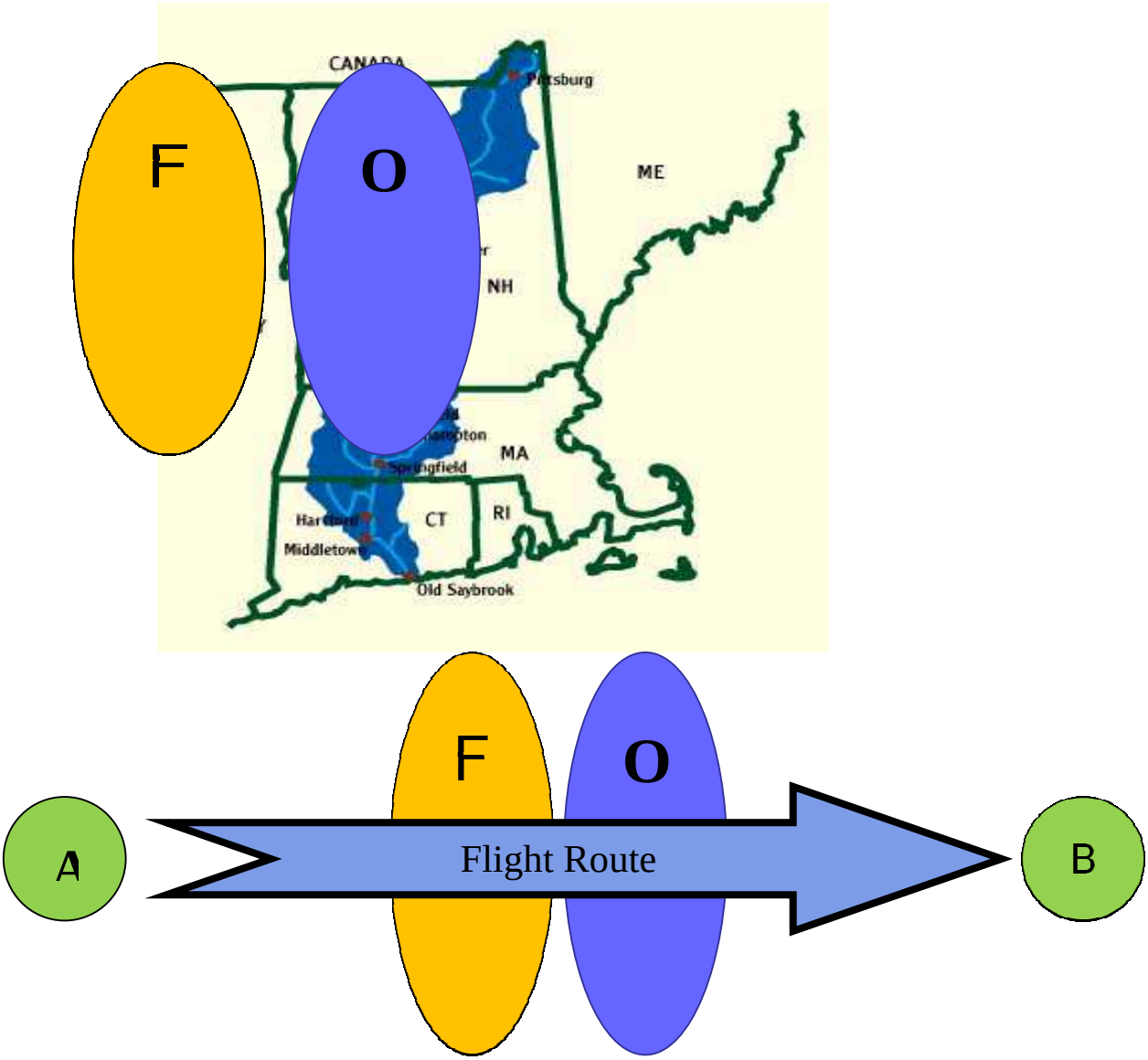

Figura 1: un esempio di come una stessa previsione possa risultare inadeguata per un'autorità di bacino (in alto) e al contempo adeguata per il gestore del traffico aereo di una compagnia aerea

naturale?
In base a questo bisogna stabilire quali “attributi” della performance previsionale e quali tecniche statistiche, di misura e grafiche adottare.
La “bontà” di una previsione è, in generale, di non semplice interpretazione (figura 1): diversi utenti possono avere idee diverse sul concetto di buona previsione, nonché diverse tecniche di verifica possono dare misure diverse di “goodness”.

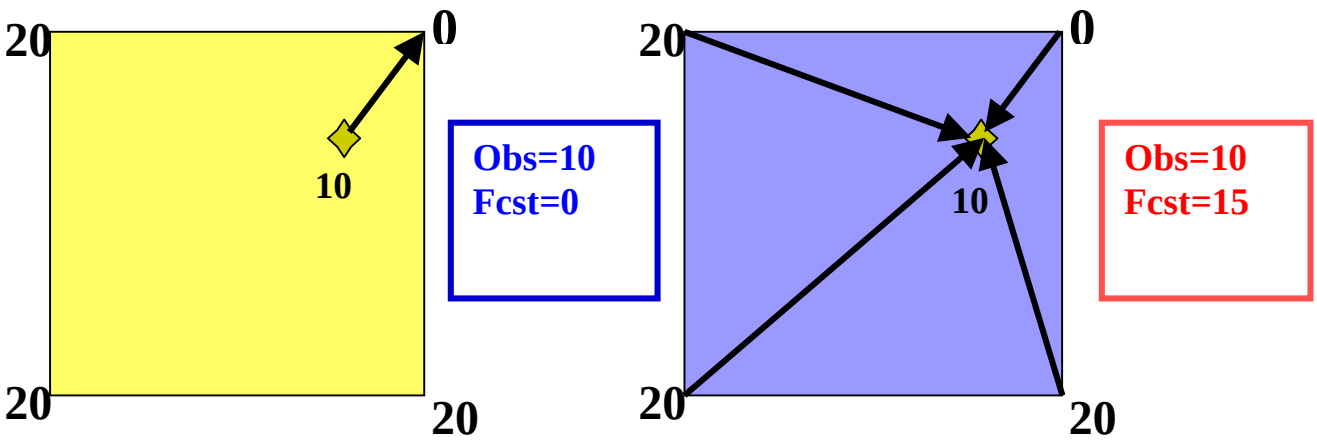

Figura 2: approccio point-to-grid (sopra) e grid-to-point (sotto). Si notino le differenze in termini di verifica “ad occhio”!

Per operare correttamente, si possono seguire delle linee guida così riassumibili:
- Considerare l’utente finale delle previsioni e della verifica: che aspetti della qualità possono essere di suo interesse?
- Porsi una domanda sull’obiettivo della verifica che soddisfi tali aspetti.
- Identificare il set di osservazioni meteorologiche disponibili per la verifica (inclusa la risoluzione spaziale, quella temporale, le variabili osservate, le soglie, le categorie, ecc.) ponendo attenzione ad eventuali stratificazioni o aggregazioni del campione che possono inficiare pesantemente il risultato della verifica.
- Identificare diversi attributi di verifica che possano rispondere alle domande (bias, correlazione, accuratezza, calibrazione, discriminazione, associazione e, per gli skill scores, proprietà – per previsioni probabilistiche - ed equitabilità – ossia previsioni random e persistenti dovrebbero avere lo stesso score).
- Selezionare misure e grafici opportuni.
- Identificare una previsione “di riferimento” per la valutazione dello skill.

L’identificazione del tipo di osservazioni e previsioni può essere così classificata: variabili **continue** (temperatura, pioggia cumulata, altezza del geopotenziale, ecc.), variabili **categoriche** (pioggia/non pioggia, vento forte/non vento forte… comunque di tipo Yes/No), variabili **multi-categoriche** (copertura nuvolosa, tipo di precipitazione, sub-set di variabili continue tipo fasce di temperatura). Per queste ultime due classi si può anche ricorrere ad una trasformazione in senso probabilistico (tipicamente il caso degli ensemble) considerando, al posto delle variabili, le loro PDF (funzioni densità di probabilità).
L’omogeneizzazione di osservazioni e previsioni per renderle vicendevolmente compatibili è l'operazione più delicata di un processo di verifica, e si articola in 2 fasi: identificazione dell’osservazione che rappresenta l’evento previsto e, in caso di dati su griglia, identificazione dell’approccio (grid-to-point o point-to-grid, vedi figura 2). Quest’ultimo procedimento comporta

variazioni nella rappresentatività di un set osservazione/previsione e può introdurre grandi differenze negli scores di verifica!
Nota bene: secondo la Brown non bisognerebbe mai usare analisi come osservazioni perché non garantiscono l'indipendenza dalle previsioni!
In termini di probabilità, un processo di verifica delle previsioni può essere rappresentato dal calcolo della probabilità congiunta (**joint probability**) di previsioni ed osservazioni p(f,o). Tale funzione può essere fattorizzata in 2 coppie di **probabilità condizionate –p(F=f|O=o)- e marginali –p(O=o):**
p(f,o) = p(F=f|O=o)p(O=o) == p (O=o|F=f)p(F=f).
Molti attributi di verifica delle previsioni derivano da queste fattorizzazioni, dette rispettivamente *"likelihood – base rate"* e *"calibration – refinement"*.
Un passo ulteriore nel processo di verifica è la scelta degli opportuni "skill scores", ossia indici che misurano la performance della previsione relativamente ad una previsione di riferimento. Essi dipendono fortemente dalla scelta del riferimento: può essere la **climatologia a priori** – climatologia effettiva ad esempio basata su 100 anni di dati -, **a posteriori** – sample climatology, basata sulle osservazioni del mio campione -, o la **persistenza**. La misura degli scores richiederebbe anche la valutazione dei **livelli di confidenza**, dal momento che nel processo di verifica emerge e si propaga l'incertezza a partire dagli errori nella classificazione del campione, dagli errori di misura delle osservabili, dalle differenze nella rappresentatività, ecc. I metodi per la valutazione dei livelli di confidenza si dividono in **metodi parametrici** (che dipendono da un modello statistico) e **non-parametrici** (che derivano da tecniche di ricampionamento o bootstrap).

Barbara Casati: verifica delle variabili continue

La verifica di variabili continue può essere effettuata graficamente, utilizzando gli scatterplot, i plot quantile-quantile (qq-plot, in cui si plottano i valori di ciascun quantile di un dataset rispetto all'altro), gli scatterplot con funzione di tabella di contingenza, i box-plot / istogrammi (distribuzioni marginali), i box-plot / istogrammi condizionati (distribuzioni condizionate), i

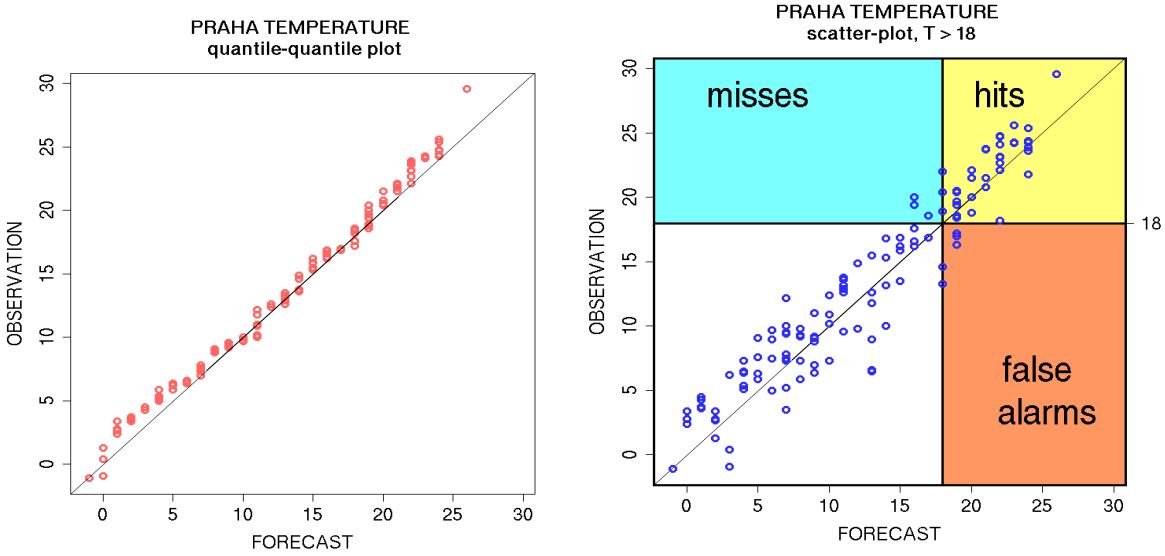

Figura 3: esempio di qq-plot (a sinistra) e scatterplot con funzione di tabella di contingenza

qq-plot condizionati.
Gli attributi che si utilizzano di solito sono: **ME** (Mean Error) ovvero il Bias (1)

(1) $$\text{linear bias} = ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) = \bar{Y} - \bar{X}$$

che dà la "direzione" della previsione (overforecast, underforecast);
MAE (Mean Absolute Error)

(2) $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

che dà una misura di accuratezza, ossia dà la "grandezza" media dell'errore nella previsione;
MSE (Mean Square Error) e **RMSE** (la sua radice)

(3) $$RMSE = \sqrt{MSE} = (\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2)^{\frac{1}{2}}$$

che dà una misura di accuratezza, ossia dà la "grandezza" dell'errore pesata sul quadrato di tutti gli errori: quest'ultimo è sensibile alla varianza specie per alte risoluzioni spaziali ma va

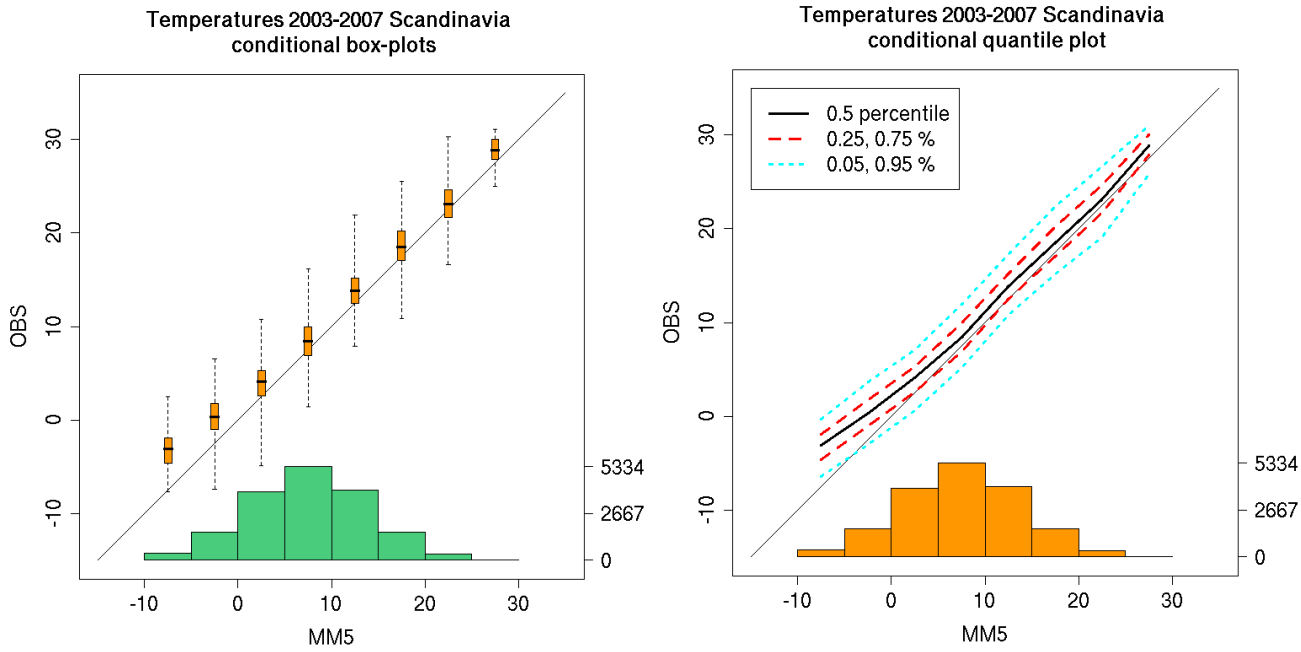

Figura 4: box-plot condizionale (a sinistra) e qq-plot condizionale

bene per ridurre i grossi contributi all'errore; il coefficiente di correlazione lineare **r**, che dà una misura di associazione lineare tra osservazioni e previsioni, ed ha il limite di scarsa robustezza (richiede in linea di principio distribuzione gaussiana) e poca resistenza (è sensibile ai grossi errori e agli outliers).
Gli skill scores che si utilizzano sono: **MAE skill score** (valore 0 per previsione perfetta)

(4) $$SS_{MAE} = \frac{MAE - MAE_{ref}}{MAE_{perf} - MAE_{ref}} = 1 - \frac{MAE}{MAE_{ref}}$$

MSE skill score (sensibile alla sample climatology ed agli estremi)

(5) $$SS_{MSE} = \frac{MSE - MSE_{ref}}{MSE_{perf} - MSE_{ref}} = 1 - \frac{MSE}{MSE_{ref}}$$

AC (Anomaly Correlation – elimina dal conto le previsioni corrette della climatologia che discendere ad esempio da effetti orografici) cumulative delle osservazioni, per cui gli errori alle code sono penalizzati rispetto agli errori intorno alla mediana e ammette tecniche di correzione della mediana che minimizzano LEPS e MAE).

$$y'_m = y_m - c_m$$
$$x'_m = x_m - c_m$$

(6)
$$AC_{cent} = \frac{\sum_{m\in map} (y'_m - \overline{y'})(x'_m - \overline{x'})}{\sqrt{\sum_{m\in map} (y'_m - \overline{y'})^2 \sum_{m\in map} (x'_m - \overline{x'})^2}}$$

$$AC_{unc} = \frac{\sum_{m\in map} (y_m - c_m)(x_m - c_m)}{\sqrt{\sum_{m\in map} (y_m - c_m)^2 \sum_{m\in map} (x_m - c_m)^2}} = \frac{\sum_{m\in map} (y'_m)(x'_m)}{\sqrt{\sum_{m\in map} (y'_m)^2 \sum_{m\in map} (x'_m)^2}}$$

scores continui di rank (utilizzano i valori di ranking della variabile anziché i valori della variabile stessa – riduce gli eventi dovuti ai valori estremi, trasforma la distribuzione marginale in uniforme, toglie il bias) e **LEPS** (Linear Error in Probability Space – è il MAE calcolato usando le frequenze

(7)
$$LEPS = \frac{1}{n}\sum_{i=1}^{n}\left|F_X(y_i) - F_X(x_i)\right|$$

Anna Ghelli: verifica di variabili categoriche

La variabile categorica rappresenta l'unica uscita di un set di possibili eventi. Tipico è il caso binario (Yes/No), relativo a previsione ed osservazione di un evento, che dà origine alla più semplice **tabella di contingenza (quella 2X2)**. In quest'ultima la probabilità marginale è data dalla somma dei valori in colonna o in riga divisa per la somma dei valori di tutte le entrate mentre la probabilità congiunta è data dall'intersezione tra una riga e una colonna. Si possono facilmente calcolare diverse misure o scores: Frequency Bias, Proportion Correct, POD, POFD, FAR, PAG (Post Agreement, definito come a/(a+b), noto anche come Frequency of Hits, in genere poco usato), CSI, ETS, PSS, HSS, OR (Odds Ratio), ORSS (Odds Ratio Skill Score). Si può fare l'estensione a più categorie per cui solo il PC può essere direttamente generalizzato, mentre gli altri scores necessitano di una riduzione ad n tabelle 2X2 del problema.

Laurie Wilson: verifica delle previsioni probabilistiche ed ensemble

Una previsione probabilistica riguarda un evento completamente definito (ad esempio la probabilità di pioggia su 6 ore). Per la verifica si utilizzano diversi scores.**Brier Score** (misura di accuracy): è di fatto l'errore quadratico medio di una previsione probabilistica, e similmente al caso deterministico assegna più peso ai grossi errori. Può essere scomposto in 3 termini che danno conto della **affidabilità** (reliability, simile al bias), **risoluzione** (resolution, capacità della previsione di distinguere

situazioni con frequenze ben distinte) e **incertezza** (uncertainty, variabilità nelle osservazioni, dipende fortemente da quanto la probabilità climatologica è vicina a 0.5); quest'ultimo termine tra l'altro impedisce di usare questo score per confrontare campioni differenti.
Brier Skill Score (misura di skill): misura il miglioramento di accuracy rispetto all'accuracy di una previsione di riferimento: la forma dello skill score si semplifica se si usa la climatologia a posteriori (sample climatology). E' sensibile ai campioni di dimensioni ridotte.
Reliability diagrams: metodo grafico per stabilire "ad occhio" la reliability (vicinanza alla diagonale), la resolution (variazione intorno alla linea orizzontale della climatologia) e lo skill (distanza dalla linea di no skill) di una previsione probabilistica. Richiede un campione ben popolato per la necessità di suddividere il campione in diversi sottocampioni condizionati alla previsione.
ROC (Relative Operating Characteristic): misura la discriminazione (discrimination, ovvero capacità di un sistema di previsione di distinguere situazioni di evento da situazioni di non evento – dipende dalla distanza dei valori medi delle distribuzioni di evento/non evento e dalla varianza

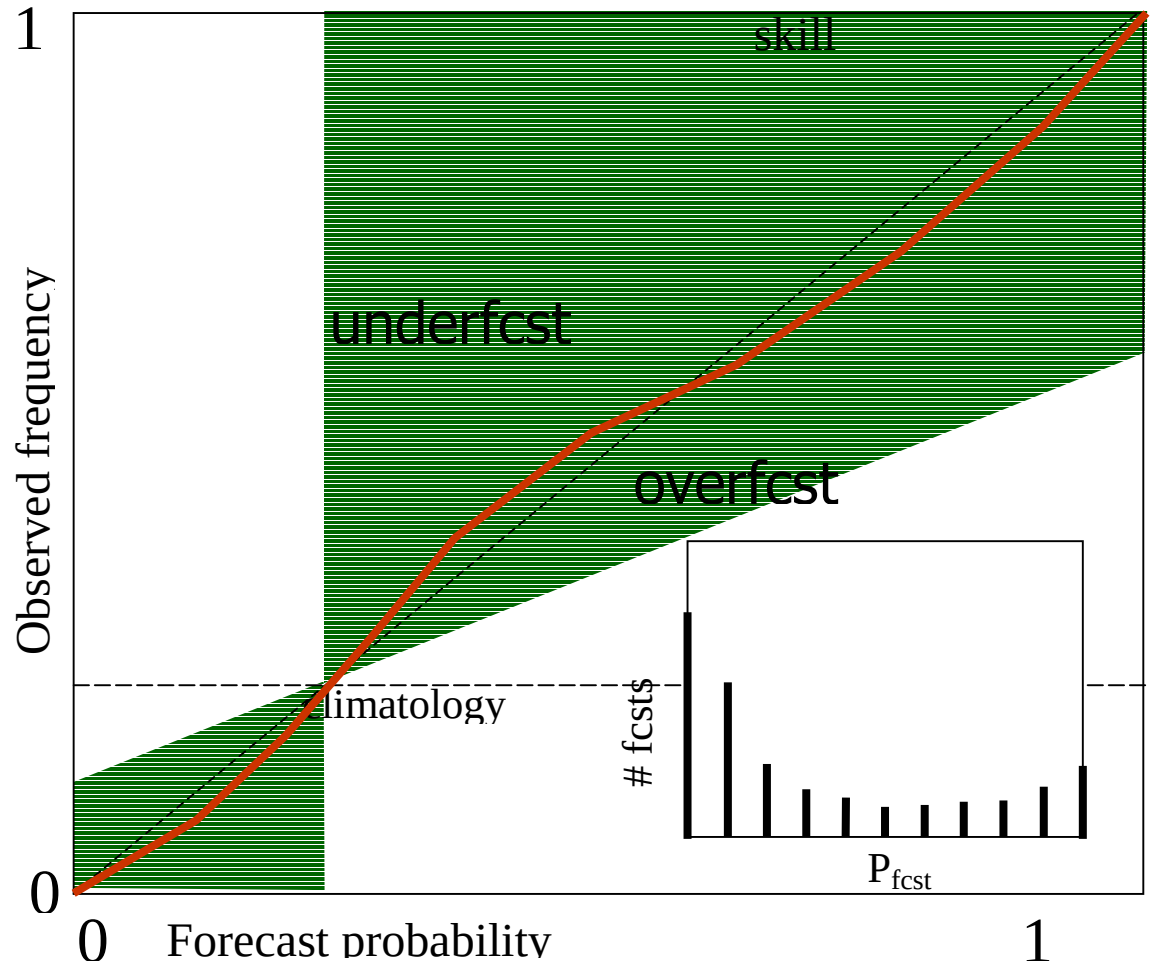

Figura 4: reliability diagram. Si notano le costanti di riferimento (climatology, skill, ecc.)

intesa come larghezza delle curve di distribuzione). Si costruisce stabilendo un numero sufficiente di bin (generalmente almeno 5) in cui suddividere il campione (difficoltà per gli eventi rari), determinando, per ogni bin, POD e POFD, plottando le coppie POD e POFD ed eventualmente usando un modello binormale per calcolare l'area sotto la curva/spezzata (qualcuno sostiene che non sia necessario scegliere un modello a priori). La curva ROC non è sensibile al bias, a meno che le due distribuzioni condizionate siano separate.

Per le tecniche di ensemble: si aggiunge una dimensione al problema poiché ci sono molti valori di previsione per una sola osservazione.
CRPS (Continuous Rank Probability Score): è uno score che tende al MAE per previsioni deterministiche, utile per gestire l'incertezza nelle osservazioni

(8) $$CRPS\ (P, x_a) = \int_{-\infty}^{\infty} \left[P(x) - P_a(x)\right]^2 dx$$

Rank Histogram: usato per analizzare lo spread medio di un ensemble in relazione alle osservazioni. Assume che le osservazioni siano equiprobabili in ciascuno degli n+1 bins.

Ian Jolliffe: intervalli di confidenza e test d'ipotesi.

Nel processo di verifica delle previsioni è necessario stimare l'incertezza nei set di osservazioni e di previsioni, in quanto usare set leggermente diversi per la medesima grandezza in linea di principio porge risultati diversi degli indici di verifica. Per fare ciò assumiamo che i dati a nostra disposizione costituiscano un campione estratto da un'ipotetica popolazione e vogliamo stabilire delle *inferenze* ad esempio riguardo a correlazione (variabili continue) e hit rate (variabili discrete) di alcune quantità (dette **parametri**: media, varianza, ecc.) di questa popolazione: facciamo la **cosiddetta analisi statistica inferenziale**.
Ci sono diversi tipi di statistica inferenziale.
Point estimation: viene considerato un singolo numero per stimare il parametro senza considerazioni sull'incertezza (la miglior stima della popolazione).
Interval estimation: si può applicare un errore standard alla miglior stima oppure calcolare un intervallo di confidenza con diverse metodologie, soluzione migliore in caso di errori distanti dall'approssimazione gaussiana.
Hypothesis testing: paragonando le stime di un parametro fatte su diversi campioni, il test d'ipotesi è un modo valido per stabilire di quanto il parametro potrebbe in generale variare.
Per quest'ultimo tipo, ci sono diversi approcci:
Inferenza parametrica classica: molto diffusa, basata sull'assunzione di distribuzioni puramente gaussiane, basata sul concetto di likelihood, *inferenza bayesiana*, *inferenza non-parametrica*, *teoria della decisione*.
Approfondimento sull'"**interval estimation**".
Il primo passo è stabilire l'intervallo di confidenza, ossia dato il valore di una misura del campione si tratta di stabilire un intervallo intorno a questo valore in cui si abbia una confidenza (espressa in percentuale) di trovare il corrispondente valore della misura nella popolazione (parametro). Nota bene: il livello di confidenza è la probabilità che l'intervallo includa il parametro, **non** la probabilità che il parametro stia nell'intervallo. E' possibile (nonché auspicabile) trovare un intervallo di confidenza anche in problemi di metrica categorica, ad esempio per l'Hit Rate: si tratta di trovare un intervallo di confidenza per la probabilità di successo p in caso di distribuzione binomiale (Yes/No). Un'approssimazione grossolana è basata sul fatto che la distribuzione di p può essere considerata gaussiana avente come media la probabilità condizionata di previsione corretta π e come varianza $p(1-p)/n$. Per valori di n piccoli possiamo assumere una distribuzione binomiale anziché gaussiana, e in questo caso gli intervalli di confidenza sono detti (impropriamente) "esatti". Un ulteriore passo avanti verso tecniche più raffinate prevede un *approccio Bayesiano* al computo degli intervalli di confidenza, che prevede la combinazione di una distribuzione "a priori" per p con una funzione di likelihood (verosimiglianza) per il dato allo scopo di ottenere una distribuzione a posteriori di p. Gli

intervalli bayesiani sono diversi dagli intervalli di confidenza, in quanto assumono che la probabilità condizionata di successo sia casuale (non fissata) e usano i percentili per la probabilità a posteriori. Nel caso di parametro distribuito binomialmente, la distribuzione di p più comunemente usata è la funzione Beta, la cui forma ha la proprietà di conservarsi anche nella probabilità a posteriori.
Un altro metodo prevede il calcolo degli *intervalli di bootstrap*: dato un campione, si considera un numero B di sottocampioni estratti a caso e di dimensione identica e si calcola p. Dopodichè si fa il ranking dei diversi p e per un intervallo di confidenza (1-2*a) si trovano il B*a-esimo più piccolo e il B*a-esimo più grande valore (detti l ed u). Ci sono diversi intervalli di confidenza di questo tipo, il più semplice dei quali prevede il calcolo dei percentili basati sulla coppia (l,u). Altri sono: basic bootstrap, parametric bootstrap, bootstrap-t intervals, ecc.

	Forecast 1	Forecast 3
Crude approx.	(0.41,0.83)	(0.78,1.03)
Better Approx.	(0.41,0.79)	(0.71,0.97)
'Exact'	(0.38,0.82)	(0.70,0.99)
Bayes – uniform	(0.41,0.79)	(0.71,0.97)
Bayes – informative	(0.48,0.79)	(0.66,0.92)
Percentile bootstrap	(0.43,0.81)	(0.76,1.00)

Figura 5: diverse stime di l e u per diversi metodi di ricerca degli intervalli di confidenza

Intervallo di confidenza sulle differenze: supponendo di avere 2 previsioni, vogliamo confrontare i rispettivi Hit Rate trovando un intervallo di confidenza per i due rispettivi parametri ($\pi 1$-$\pi 2$). E' in generale più appropriato trovare l'intervallo di confidenza della differenza piuttosto che l'intersezione dei singoli intervalli, che spesso è fuorviante. In generale le tecniche di bootstrap per il calcolo degli intervalli di confidenza vanno usate quando si hanno dubbi sulle distribuzioni sottostanti.
Approfondimento sull'"**Hypothesis testing**".
L'interesse nella valutazione dell'incertezza associata con una misura di verifica porta a domande del tipo: il valore osservato è compatibile con quanto si osserverebbe se il sistema di previsione avesse skill nullo? Oppure: date due misure provenienti da due sistemi diversi di previsione (o dallo stesso sistema in tempi diversi), è possibile che le differenze nei valori nascano casualmente, supponendo che non vi siano differenze nello skill dei due sistemi (o nei due tempi)? La risposta chiara viene dal test d'ipotesi nulla di "no skill" nel primo caso o di "skill identico" nel secondo caso: si rigetta l'ipotesi nulla quando il valore cade al di fuori del prediction interval del 95%. I test d'ipotesi possono esser trattati come processi decisionali sulla base del livello di significatività (5%, 1%, ecc.) assegnato. Per una teoria della decisione completa, è necessario dotarsi di una *loss function* e della *prior probability*, o in alternativa di un *p-value*, ossia della probabilità che il dato possa essere trovato casualmente se l'ipotesi nulla è vera (nota bene: non è la probabilità che l'ipotesi nulla sia vera!).
Se non siamo pronti a fare assunzioni sulla distribuzione della variabile, possiamo usare un approccio di permutazioni: fissato ciascun valore delle previsioni, si considerano tutte le

permutazioni delle corrispondenti osservazioni e si calcolano le correlazioni tra le previsioni e le osservazioni in tutti i casi permutati (sotto il test di ipotesi nulla tutte le permutazioni hanno la medesima probabilità). Se il numero delle permutazioni è molto grande (e di fatto accade sempre così), si usa un test di randomizzazione, ossia si seleziona un numero random di subset.

Beth Ebert: metodi per la verifica di previsioni spaziali.

Le previsioni di variabili spaziali (ovvero di campi scalari) richiedono delle tecniche particolari di verifica. Il primo problema è trovare la corretta forma per valutare la sovrapponibilità delle previsioni e delle osservazioni: si può scegliere un approccio point-to-grid o grid-to-point, ma come si è scritto in precedenza esso non garantisce la conservazione dei risultati della verifica; oppure si può cercare la sovrapponibilità di grigliati regolari di osservazioni e previsioni. In ogni caso, è necessario avere un sistema di osservazioni diffuso nello spazio, a risoluzione paragonabile alle previsioni (es. radar QPE – Quantitative Precipitation Estimate).

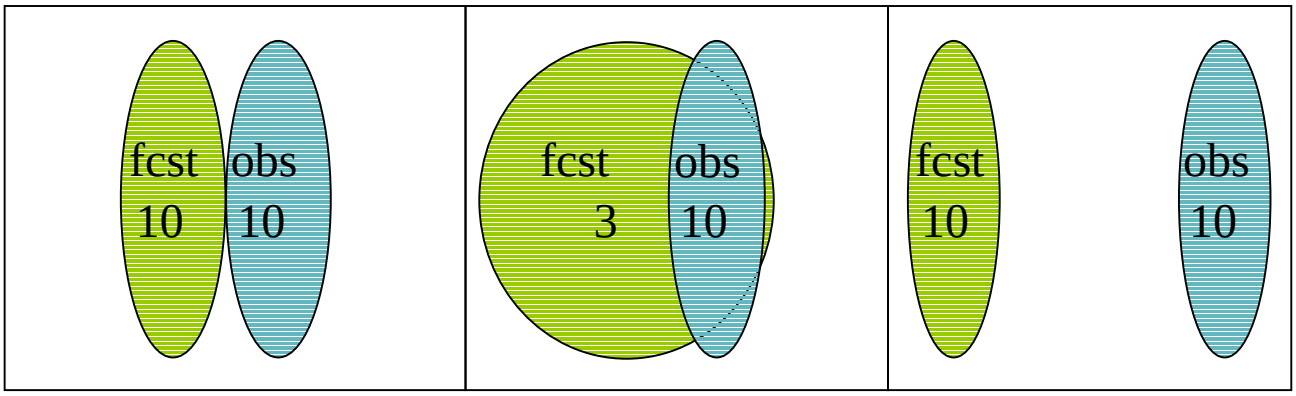

Figura 6: modelli ad alta risoluzione (sinistra e destra) e modelli a bassa risoluzione (al centro).

Gli approcci tradizionali alla verifica di previsioni estese nello spazio possono trattare variabili continue (es. la quantità di pioggia caduta) o categoriche (es. l'evento di pioggia Yes/No), con i ben noti indici. Il limite principale è che essi richiedono un'esatta sovrapposizione di osservazioni e previsioni in ogni punto di griglia, e per modelli ad alta risoluzione ciò è molto penalizzante (vedi figura 6). Inoltre i metodi tradizionali non dicono molto della natura dell'errore di previsione, oltre a non rendere l'idea di quanto una previsione sia realistica o possa essere d'aiuto nel processo di "decision making". Paradossalmente valorizzano le previsioni "smooth" e rimangono in molti casi insensibili alle dimensioni dell'errore. Considerando il fatto che i campi scalari meteorologici mantengono generalmente una loro struttura spaziale coerente, sono state sviluppate nuove tecniche che tengano conto di ciò, oltre a fornire informazioni sulla natura fisica dell'errore e sulle incertezze nello spazio e nel tempo. Queste nuove tecniche utilizzano diversi approcci: **metodi "fuzzy"** (danno valore alle previsioni "vicine" alle osservazioni), **metodi di decomposizione di scala** (misurano gli errori dipendenti dalla scala), **metodi "object-oriented"** (valutano gli attributi propri di oggetti identificabili), **verifica di campi** (valutano gli errori di fase).

Metodi "fuzzy" (neighborhood): non richiedono un'esatta sovrapponibilità tra previsioni ed osservazioni, prendendo in considerazione ciò che accade nelle vicinanze (sia di spazio che di tempo) del punto di interesse. Generalmente si prende un sottodominio (detto "finestra") in cui si calcola il valor medio della grandezza in oggetto, oppure la variabile categorica Yes/No di un

evento in uno o più punti della finestra, anche in senso probabilistico, o la distribuzione dei valori nella finestra. Si procede poi alla definizione di diversi set di sottodomini, a risoluzione spaziale (o temporale) via via decrescente (***upscaling***), utilizzando le consuete metriche di verifica: MSE, r, RMSE in caso di variabili continue, scores e attributi della tabella di contingenza in caso di variabili categoriche. Una tecnica più moderna si basa sulla definizione del ***Fractions Skill Score***

$$\text{FSS} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(P_{fcst} - P_{obs})^2}{\frac{1}{N}\sum_{i=1}^{N}P_{fcst}^{\,2} + \frac{1}{N}\sum_{i=1}^{N}P_{obs}^{\,2}} \qquad (9)$$

che dà indicazione sulla variazione dello skill del sistema di previsione a diverse risoluzioni spaziali, dando indicazione della più alta risoluzione a cui lo skill è accettabile o significativo (ad esempio verifica lo skill di un sistema NWP ad alta risoluzione a scala di bacino per utilizzi idrogeologici): in pratica, mette in relazione previsioni ed osservazioni in senso probabilistico. Un'estensione dei metodi fuzzy è anche la tabella di contingenza per ***eventi multipli nello spazio***, basata sulla costruzione di una curva ROC, che consente di valutare se un sistema di previsioni distingue bene tra evento e non evento (utilizzando ad esempio uno score come il PSS) oltre una certa soglia di decisione e ad una certa risoluzione spaziale.

Metodi di decomposizione di scala: valuta lo skill di un sistema di previsione come funzione dell'intensità e della scala spaziale dell'evento. Sfruttando delle soglie di intensità si producono variabili binarie di osservazione e previsione che vengono decomposte a scale via via crescenti secondo la tecnica della "wavelet decomposition", calcolando uno skill score sull'errore quadratico medio (ad esempio il metodo "intensity-scale" sviluppato da Barbara Casati).

Metodi orientati agli oggetti meteorologici estesi nello spazio: il più noto è il metodo CRA (***Contiguous Rain Area***) che definisce un'entità meteorologica, estesa nello spazio, basata su soglie (ad esempio, un'area di precipitazione) e quantifica la traslazione orizzontale necessaria a soddisfare un criterio di "pattern matching" predefinito: minimo errore quadratico totale, massima correlazione, massima sovrapponibilità tra previsione ed osservazione. Il cosiddetto "spostamento" è la differenza vettoriale tra la posizione iniziale e la posizione finale dell'oggetto in questione. Il MSE totale che si calcola con questo metodo è scomponibile in 3 termini: l'errore di *spostamento* (differenza tra MSE prima e dopo lo spostamento), l'errore di *volume* (il bias nell'intensità media) e l'errore di *pattern* (residuale, caratteristico di differenze nelle strutture più fini).

Metodo per la valutazione diagnostica basata sull'oggetto (MODE): decompone gli oggetti meteorologici secondo alcune proprietà (posizione del centroide, distribuzione delle intensità, area, orientazione, ecc.) e calcola 2 parametri: il raggio di convoluzione e le soglie.

Structure Amplitude Location (SAL): dato un dominio e una soglia di precipitazione, calcola gli errori in ampiezza, in posizione ed in struttura.

Field verification – errori di fase: uno dei metodi è il **DAS** (Displacement and Amplitude Score) che combina distanza ed

ampiezza portando previsioni su osservazioni e osservazioni su previsioni.
In conclusione ogni domanda di verifica ha la sua risposta in termini di metodo appropriato, anche in assenza di alcuni dei prerequisiti (es. il campo spazializzato ad alta risoluzione tipico della Radar QPE). Ad esempio, verifiche di campi di vento aventi a disposizione solo osservazioni da gauges (caso tipico in FVG), possono essere soddisfatte grazie ai metodi fuzzy. Per ulteriori approfondimenti, vedi lo specchietto in calce a questo tutorial.

Sessione dei working gropus

A margine delle lezioni del tutorial sono stati organizzati dei lavori di gruppo in cui sono stati affidati agli studenti dei dataset (per lo più portati dagli studenti stessi) che richiedevano diversi problemi di verifica: dalla QPF di modelli ad alta risoluzione alle previsioni di vento negli aeroporti, dalla climatologia dei cicloni tropicali alla verifica dei radiosondaggi. Il mio gruppo ha avuto in dote un dataset contenente le previsioni della pioggia in Catalogna da modello MM5 a risoluzione spaziale 12 km, da verificare con la Radar QPE estratta dalla rete radar catalana. La tipologia di dataset ha richiesto l'utilizzo di tecniche spaziali, e noi abbiamo scelto di esplorare le tecniche fuzzy; gli indici tradizionali della metrica categorica non hanno fornito risultati incoraggianti, mentre il Fractions Skill Score ha invece dimostrato che per risoluzione spaziale un po' più grossolana di quella nativa del modello, e per soglie di pioggia piccole ma non piccolissime (sotto i 5 mm in 6 ore) si ha una previsione accettabile (FSS più alto di 0.5).
E' quindi un tipico esempio in cui cercando lo score più adatto si riesce ad estrarre un valore (non in senso economico!) da una previsione che, apparentemente o per diversi set di verifica, non ne ha.

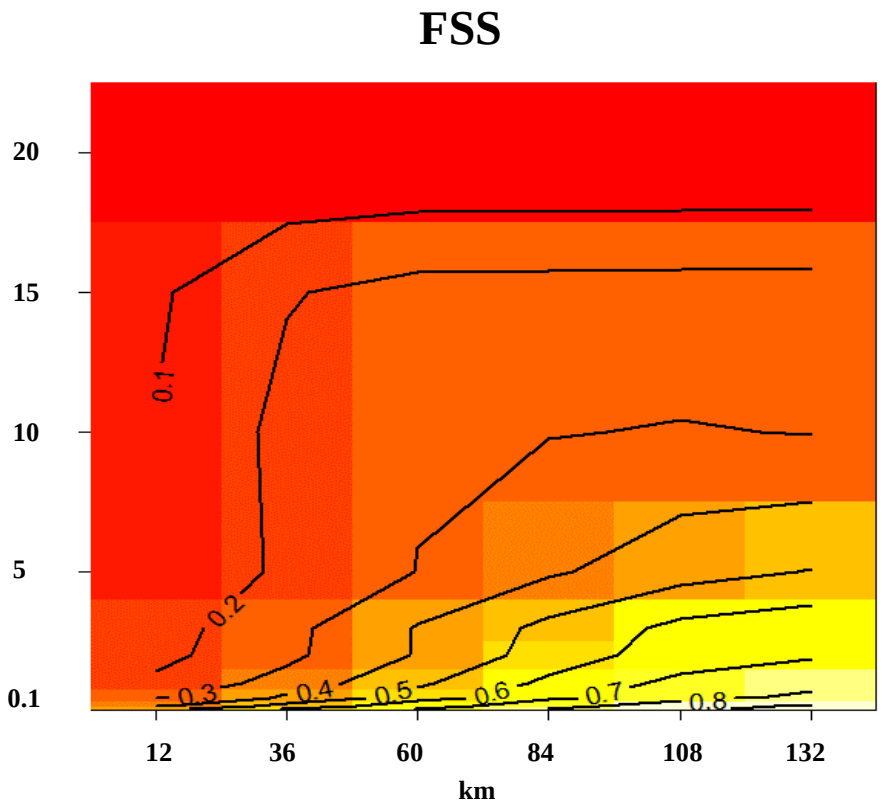

Figura 7: diagramma del Fractions Skill Score. Il valore 0.5 si intende come valore discriminante

Bibliografia

Jolliffe and Stephenson (2003): Forecast verification: A practitioner's guide, Wiley & sons
Wilks (2005): Statistical Methods in Atmospheric Science, Academic press, 467 pp.
Efron B and Tibshirani RJ (1993). An Introduction to the Bootstrap. New York: Chapman & Hall.

Epstein ES (1985). Statistical Inference and Prediction in Climatology: A Bayesian Approach. Meteorological Monograph. American Meteorological Society.
Jolliffe IT (2007). Uncertainty and inference for verification measures. Wea. Forecasting, 22, 637-650.
Tuyl F, Gerlach R & Mengersen K (2008). A comparison of Bayes-Laplace, Jeffreys, and other priors: the case of zero events. Amer. Statist., 62, 40-44.

Nella tabella qui sotto un breve schema per utilizzare le forme appropriate di verifica di tipo spaziale con relativa bibliografia (vedi presentazione di Beth Ebert).
NO – NF = neighborhood observation / neighborhood forecast
SO – NF = single observation / neighborhood forecast

Neighborhood method	*Matching strategy**	*Decision model for useful forecast*
Upscaling ***(Zepeda-Arce et al. 2000; Weygandt et al. 2004)***	***NO-NF***	***Resembles obs when averaged to coarser scales***
Minimum coverage ***(Damrath 2004)***	***NO-NF***	***Predicts event over minimum fraction of region***
Fuzzy logic ***(Damrath 2004), joint probability (Ebert 2002)***	***NO-NF***	***More correct than incorrect***
Fractions skill score ***(Roberts and Lean 2008)***	***NO-NF***	***Similar frequency of forecast and observed events***
Area-related RMSE ***(Rezacova et al. 2006)***	***NO-NF***	***Similar intensity distribution as observed***
Pragmatic ***(Theis et al. 2005)***	***SO-NF***	***Can distinguish events and non-events***
CSRR ***(Germann and Zawadzki 2004)***	***SO-NF***	***High probability of matching observed value***
Multi-event contingency table ***(Atger 2001)***	***SO-NF***	***Predicts at least one event close to observed event***
Practically perfect hindcast ***(Brooks et al. 1998)***	***SO-NF***	***Resembles forecast based on perfect knowledge of observations***