Verificación probabilista

15

DOI: 10.31978/014-18-009-X.15

CARLOS SANTOS BURGUETE
Centro Nacional de Predicción (CNP), Agencia Estatal de Meteorología (AEMET)

CONTRIBUCIONES DE: M. ASUNCIÓN PASTOR SAAVEDRA
Área de Modelización y Evaluación del Clima, AEMET

La evaluación es una de las dimensiones fundamentales de la educación y también de otros campos, aunque la atención que se le dedica no es proporcional a su importancia.

ISABEL CAPPELLETTI, profesora universitaria y pedagoga brasileña

Los modelos de la atmósfera son simulaciones mediante supercomputación de los procesos físicos que tienen lugar en la misma. Estos modelos han mejorado enormemente en las últimas décadas, revolucionando el mundo de la predicción, debido a los avances también tremendos en tecnología de la información, sistemas de observación por satélite, técnicas de asimilación de datos y en la formulación del propio modelo. ¿Cómo se evalúan estos avances? ¿Son perfectos los modelos? ¿Cómo se mide la calidad de los modelos? ¿Qué nos permite decir si un modelo es más adecuado que otro según para qué propósitos?

Palabras clave: verificación de modelos atmosféricos, verificación de sistemas de predicción por conjuntos, métodos de verificación estadísticos, métodos de verificación espaciales, verificación probabilista, valor económico relativo.

Imagen parte superior: *cumulus congestus* (Cu con). Usanos (Guadalajara), hacia el noreste, 18 de mayo de 2007, a las 19:55. Fotografía de JOSÉ ANTONIO QUIRANTES CALVO.

15.1 Introducción a la verificación

Los modelos de predicción numérica del tiempo son simulaciones de la atmósfera y sus interacciones. Para medir su calidad y valor [35] deben ser comparados con una buena representación de la realidad, que pueden ser las observaciones mismas o, en su lugar, los análisis correspondientes. Este proceso de comparación lo denominaremos verificación, aunque en otros contextos puede llamarse validación o evaluación. En ese proceso usaremos unas métricas o medidas que denominaremos *scores* (se utiliza el término anglosajón por su universalidad). Se ayuda así a los modelizadores a mejorar los modelos y como guía de uso adecuado de los mismos para los predictores del tiempo. En particular, podremos medir la calidad, la exactitud y las tendencias del modelo; comprender sus fallos; comparar diferentes modelos o sistemas de predicción (e. g. comparar la versión operacional del modelo con algunas versiones experimentales). Se suele distinguir entre verificación administrativa, económica y científica. En todas ellas, la verificación es una eslabón importante en la toma de decisiones. Las observaciones y las predicciones pueden compararse en diferentes marcos: como conjuntos que tienen ciertas propiedades estadísticas, o bien por parejas y luego estadísticamente resumir las comparaciones, o bien como distribuciones, etc. Los métodos de verificación estandarizados se describen con detalle en [48] y en [26]. La verificación de SPC se aborda con detalle en [7] o en [43].

15.1.1 Calidad, valor y consistencia

La calidad de un modelo está relacionada con la correspondencia entre observación y predicción, mientras que el valor está asociado a la utilidad para el usuario a la hora de tomar decisiones. Un ejemplo que ilustra perfectamente esta diferencia es el siguiente: una predicción de cielos despejados para el Desierto del Sáhara goza es de mucha calidad, pero con poco valor al no aportar información relevante; sin embargo, una predicción de una probabilidad de 30 % de que los vientos superen los 100 km/h en un temporal en el Cantábrico, no tiene la calidad de la primera, pero será de gran valor en la toma de decisiones, aportando información relevante. Puede hablarse también de consistencia entre modelo y juicio del predictor.

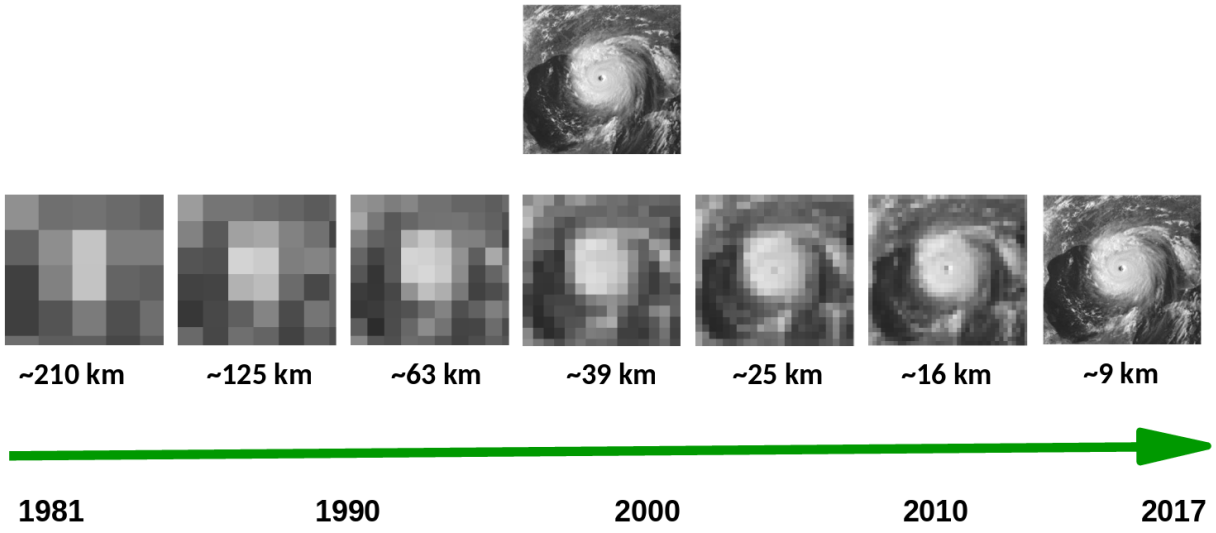

Figura 15.1: Los modelos atmosféricos han mejorado enormemente en las últimas décadas. Aquí, una verificación subjetiva muestra un huracán observado (arriba) comparado visualmente con diferentes épocas del modelo que lo simula a lo largo de la historia, con diferentes resoluciones debido al avance científico y tecnológico (abajo). Figura adaptada de ECMWF.

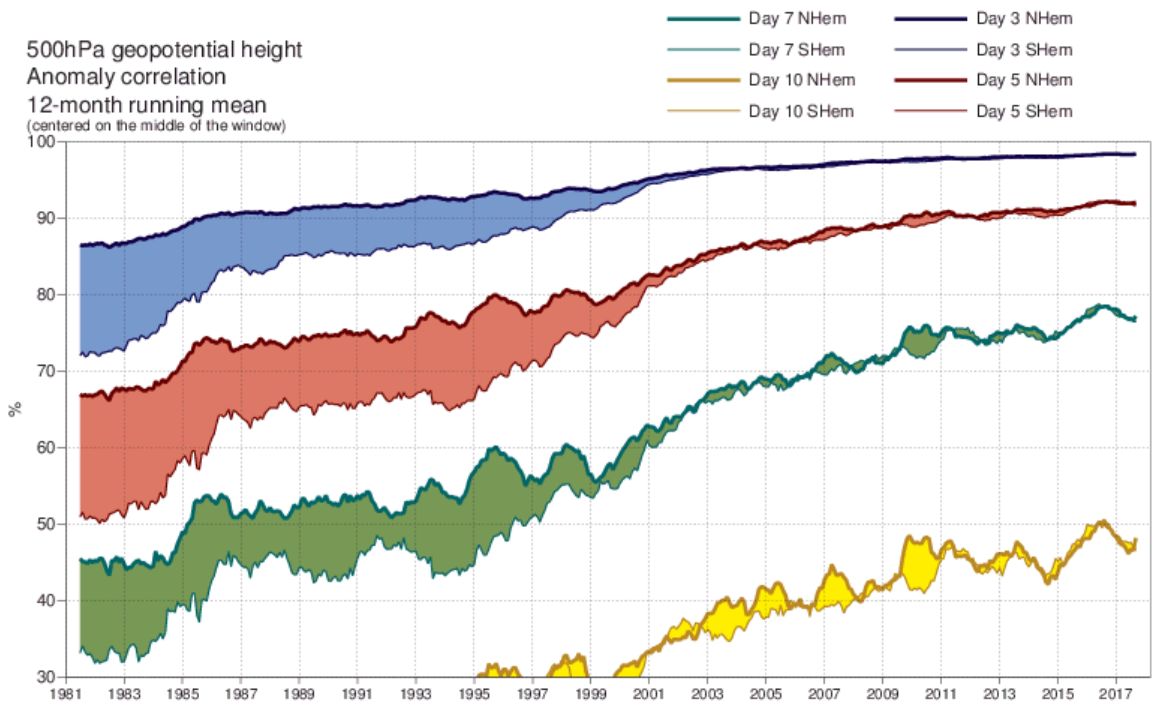

Figura 15.2: Los modelos atmosféricos han mejorado enormemente en las últimas décadas. Aquí una verificación objetiva muestra la mejora a largo plazo en la calidad del mismo modelo que en la Figura 15.1 en la página anterior pero con medidas cuantitativas, en este caso el coeficiente de correlación de anomalías (ver texto) representado en un gráfico de serie temporal. ECMWF.

15.1.2 Subjetiva, objetiva y diagnóstica

La *verificación subjetiva* es la más inmediata, se basa en estimaciones cualitativas, en un caso óptimo, de especialistas en predicción o en climatología, a menudo basada en la mera comparación visual entre modelo y observaciones, a veces basada en la experiencia; brinda información llamada *diagnóstica* porque resalta puntos débiles y fuertes del modelo en términos de su comportamiento meteorológico. Por otro lado, la *verificación objetiva* se basa en información cuantitativa, supuestamente con significación estadística, no siempre detallada y no siempre explícita; no brinda información diagnóstica pues no informa sobre puntos fuertes o débiles del modelo en términos meteorológicos puros.

Para ilustrar la diferencia tomemos como ejemplo el modelo determinista del European Centre for Medium-range Weather Forecasts -Centro Europeo de Predicción a Plazo Medio- (ECMWF), denominado ECHRES (sec. 19.2 en la página 291).

La Figura 15.1 en la página anterior muestra el aumento en resolución horizontal del modelo ECHRES a lo largo de su historia: la precisión con la que pueden describirse fenómenos atmosféricos como un huracán está íntimamente relacionada con la resolución horizontal (así como la vertical y la temporal) del modelo.

Aunque no se proporcione una medida cuantitativa de la mejora, resulta evidente inspeccionando la figura. Es un ejemplo de verificación subjetiva.

Por otro lado, en la Figura 15.2 podemos apreciar una verificación objetiva donde se muestra gráficamente la mejora del mismo modelo ECHRES, medida en este caso de forma cuantitativa (pueden consultarse detalles en el apartado 15.2 en la página siguiente) con el índice de correlación entre anomalías, con respecto a la climatología, de las predicciones y de las observaciones; variando entre 0 y 100, cuanto más alto, mas calidad tiene el modelo. Las curvas tienden a crecer según avanzan los años: como tendencia general la calidad del modelo va mejorando.

Por último, la *verificación diagnóstica* es aquella que, a caballo entre subjetiva y objetiva, brinda información basada en medidas cuantitativas objetivamente fundamentadas, pero a la vez nos informa sobre aspectos fuertes y débiles del modelo en términos puramente meteorológicos. Ejemplos de verificación diagnóstica son *el modelo adelantó la precipitación tres horas* o *el modelo falló en la colocación del sistema convectivo por 50 km*. La verificación diagnóstica no fue más que un sueño hasta que a principios del siglo XXI se empezaron a aplicar sistemáticamente los métodos espaciales (ver apartado 15.5 en la página 219).

Verificación frente a análisis	Verificación frente a observaciones
Favorece al modelo Comparación endogámica Cada análisis es diferente	Más exigente Problemas de cobertura Escalas espaciales de los modelos y observaciones Error en las observaciones Representatividad

Tabla 15.1: Verificación frente a observaciones vs verificación frente a los análisis.

15.1.3 Frente a observaciones y frente a análisis

Como ya se dijo antes, podemos comparar el modelo directamente con las observaciones (por abuso del lenguaje se suele decir "contra las observaciones") o con una buena representación de la realidad como puede ser el análisis del modelo (ver tema sobre asimilación; un análisis es el resultado de adaptar las observaciones a la grilla del modelo obligando así mismo a una serie de restricciones físicas/meteorológicas a las mismas y a un cierto "control de calidad", por lo cual el análisis puede ser considerado "algo más suave que las observaciones puras"). Ambas comparaciones tienen sus virtudes y sus defectos, que enumeramos a continuación.

Naturaleza del predictando	Ejemplos	Métodos estadísticos
Continua	Temperatura	Diagrama de dispersión, BIAS, MSE...
Dicotómica	Sí llueve / No llueve	TSS, FBI...
Multicategórica	Cálido / Normal / Frío	Distribuciones, HSS...

Tabla 15.2: Versión sencilla de la clasificación de MURPHY *de las verificaciones por naturaleza de la variable.*

15.1.4 Tipos de variables

Las diferentes variables meteorológicas de interés, por su distinta naturaleza, requieren métodos de verificación también diferentes. Una primera clasificación sencilla siguiendo este criterio [28, 33, 34, 48] considera tres grupos de variables: *continuas*, *dicotómicas* y *multicategóricas*. Las variables continuas pueden tomar cualquier valor dentro de su rango natural de variabilidad, las dicotómicas son binarias tipo «llueve o no llueve». Por último, las multicategóricas admiten más categorías; por ejemplo las predicciones estacionales proporcionan el carácter de la estación con respecto a la climatología (cálido / normal / frío). La tabla 15.2 ilustra esta clasificación. Los métodos específicos de cada grupo pueden consultarse en el apartado 15.2.

15.2 Métodos estadísticos clásicos

Desde el punto de vista estadístico, las observaciones y las predicciones pueden compararse, bien en conjunto, utilizando las propiedades estadísticas de cada conjunto, o bien por parejas {observación, predicción} (abreviadamente {o, p}) y estudiando las propiedades estadísticas del conjunto de parejas.

15.2.1 Métodos descriptivos, orientados a scores y orientados a distribuciones

Usamos diagramas descriptivos para dar un primer paso, una idea general y con perspectiva de los conjuntos de datos previstos y observados. Si además usamos estimadores estadísticos que resuman estos comportamientos, tendremos la verificación orientada a medidas o scores. La verificación orientada a distribuciones proporciona información más rica por no tratarse de un resumen.

15.2.2 Métodos descriptivos

Diagrama de dispersión (scatter plot). Los puntos representan predicción vs observación dando una idea muy visual de la correspondencia entre predicciones y observaciones. En un modelo con mucha exactitud cada predicción será muy similar a la correspondiente observación, de modo que presentará puntos cerca de la diagonal. Las desviaciones de la diagonal nos informan sobre los errores del modelo, en diferentes rangos de valores, etc.

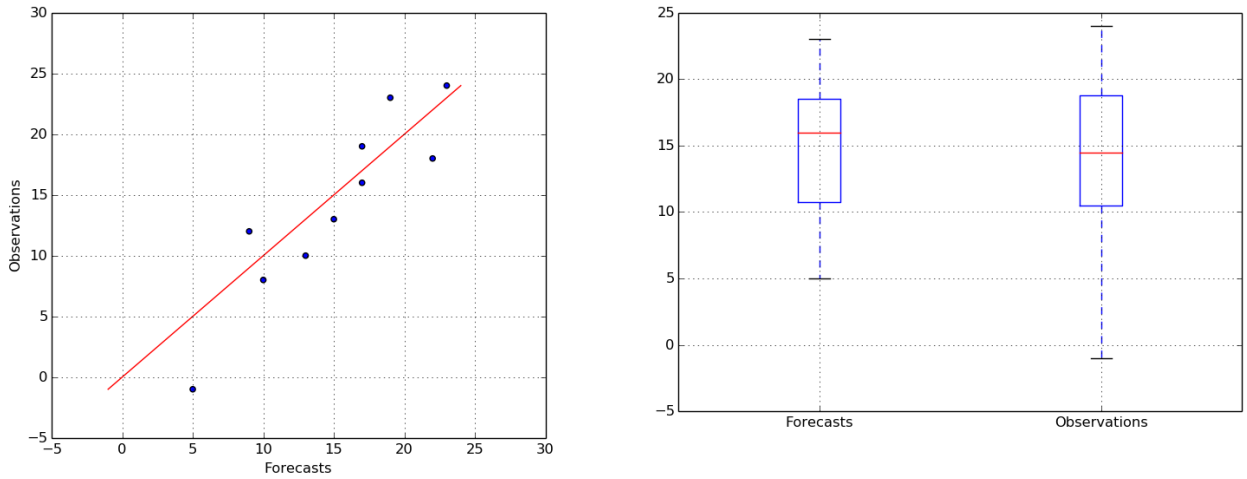

Figura 15.3: Métodos descriptivos: diagrama de dispersión (scatter plot) a la izquierda y diagrama de caja (box-plot) a la derecha.

Diagrama de caja (box plot). Muestra en una caja información condensada y visual sobre la distribución, bien sea de observaciones, bien de predicciones. Dentro de la caja cae el *rango intercuartílico* (sus bordes superior e inferior son los percentiles 25 y 75). La raya horizontal dentro de la caja muestra la mediana (no la media), y los bigotes (*whiskers*) superior e inferior muestran el máximo y el mínimo, es decir, el rango completo de variación de la distribución. Muestra información sobre la correspondencia de la distribución de predicciones con la distribución de observaciones. No brinda información sobre la correspondencia entre predicciones y observaciones. Nos sirve para comparar propiedades típicas como la localización, la dispersión y la asimetría de las distribuciones de predicciones y de observaciones.

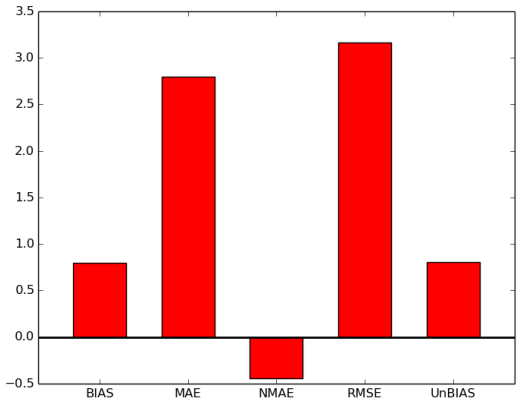

Figura 15.4: Métodos orientados a medidas (scores): histograma con medidas usuales.

15.2.3 Medidas o scores y atributos de calidad

Los atributos de calidad están relacionados con diferentes aspectos de la calidad del modelo. Algunos de ellos van asociados a un *score* o a varios, otros se relacionarán con una idea no siempre cuantificable.

Imaginemos el siguiente ejemplo. Hagamos un pronóstico de temperatura para mañana a las 00 UTC en Sevilla: 4º C. Al día siguiente recogemos el dato observado: 3º C. A la pareja formada por la predicción y la observación la podemos denotar por $\{4, 3\}$. La medida de calidad más elemental es el error. Error del modelo = predicción – observación = 1º C. ¿Qué información podemos inferir a partir de esta verificación? Se invita al lector a reflexionar este punto. ¿Puede decirse realmente que el modelo tiene calidad? ¿Por qué? Lo único que podemos realmente afirmar es que el modelo tuvo cierta calidad (un error de 1º C), para ese día, a esa hora, en Sevilla. Pero nada podemos afirmar sobre su calidad otros días, otras horas o para otras localidades. Para dar respuestas a esas preguntas hemos de tomar muestras más grandes de datos, y dar resultados estadísticos con esas muestras. Por ejemplo, tomando parejas de predicciones y observaciones durante todos los días del invierno en Sevilla, haremos el promedio de los errores, al que llamaremos error medio. En esta sección ofrecemos una relación, no exhaustiva pero sí precisa, de scores o medidas de calidad, abarcando una buena parte de los atributos de calidad más importantes. Mantendremos a menudo, por su utilidad, el término en inglés acompañando al teŕmino en castellano; en ocasiones es preferible el

uso del término anglosajón para evitar confusiones entre diferentes fuentes bibliográficas.

Error medio (bias, mean error). Promedio entre las diferencias entre predicción y observación. Tiene problemas de compensación: en una muestra puede haber partes con error medio positivo y otras con error medio negativo y pueden compensarse dando un error medio casi nulo que no refleja esos errores parciales. El rango de variación del error es $(-\infty,+\infty)$ y una predicción perfecta tiene error 0.

$$\frac{1}{N}\sum_{i=1}^{N}(p_i-o_i) \tag{15.1}$$

Error absoluto medio. Promedio entre las diferencias absolutas entre predicción y observación. Evita los problemas de compensaciones del error medio. El rango de variación es $[0,+\infty)$ y una predicción perfecta tiene 0.

$$\frac{1}{N}\sum_{i=1}^{N}|p_i-o_i| \tag{15.2}$$

Error cuadrático medio (mean square error). Promedio de los cuadrados de las diferencias entre predicciones y observaciones individuales. Evita los problemas de compensaciones del error medio, pero penaliza mucho errores grandes. El rango de variación es $[0,+\infty)$ y una predicción perfecta tiene 0.

$$\frac{1}{N}\sum_{i=1}^{N}(p_i-o_i)^2 \tag{15.3}$$

Raíz del error cuadrático medio. Sus unidades coinciden con las de la magnitud. El rango de variación es $[0,+\infty)$ y una predicción perfecta tiene 0.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i-o_i)^2} \tag{15.4}$$

Coeficiente de correlación de anomalías. Correspondencia entre anomalías respecto de la climatología, en forma de coeficiente de correlación. El rango de variación es $[-1,+1]$ y una predicción perfecta tiene 1.

$$\frac{\sum_{i=1}^{N}(p_i-c_i)\sum_{i=1}^{N}(o-c_i)}{\sqrt{\sum_{i=1}^{N}(p_i-c_i)^2}\sqrt{\sum_{i=1}^{N}(o_i-c_i)^2}} \tag{15.5}$$

La Tabla 15.3 en la página siguiente ofrece un resumen de las medidas básicas más relevantes relacionadas con atributos de calidad, con su interpretación, su rango de variación y el score que obtendría una predicción perfecta.

Para las diferentes definiciones de la tabla 15.3 en la página siguiente se denota el promedio para todos los datos con el operador $\overline{()}=\frac{1}{N}\sum_{i=1}^{N}()$, considerando que hay N observaciones o_i y N predicciones p_i, es decir, N parejas $\{p_i,o_i\}$; los valores climatológicos son c_i. En ese sentido tendremos, por ejemplo, que $\overline{p-o}=\frac{1}{N}\sum_{i=1}^{N}(p_i-o_i)$

15.2.4 Skill scores

Habilidad, destreza o pericia (*skill*) tiene que ver con la exactitud relativa de la predicción con respecto a una predicción de referencia. Esta predicción de referencia suele tomarse sin habilidad, tal como una predicción aleatoria, la persistencia (última observación disponible, prolongada) o la climatología. El skill se refiere a la mejora en exactitud debida a la habilidad del modelo. A veces las predicciones son más exactas sencillamente porque en esas ocasiones el tiempo es más fácil de predecir y el skill tiene esto en cuenta. Los así llamados skill scores o medidas de habilidad surgen, por tanto, de forma natural para poder comparar la calidad del modelo con la calidad de un sistema predictivo de referencia. Expresado de otro modo, ¿qué mejora relativa me aporta este modelo con respecto a una referencia?

$$skill\ score=\frac{score_{predicción}-score_{referencia}}{score_{predicción\ perfecta}-score_{referencia}} \tag{15.6}$$

En lo referente al rango de variación de un skill score podemos decir que el límite inferior depende, naturalmente, del score en cuestión y del sistema predictivo de referencia usado; el límite superior es siempre 1, mientras que 0 indica la ausencia de mejora sobre el sistema de referencia. En Meteorología, el sistema predictivo de referencia suele ser la persistencia (sin cambios desde la última observación) o la climatología muestral (las mismísimas observaciones de la muestra). El skill score puede ser inestable para muestras pequeñas. Cuando el score utilizado es el error cuadrático medio (MSE) entonces el skill score se llama reducción de la varianza.

Medida (score)	Definición	Rango	Perfecto
Error medio (bias, mean error). Promedio entre las diferencias entre predicción y observación, opuesto a la *exactitud*. Tiene problemas de *compensación*: en una muestra puede haber partes con error positivo y otras con negativo, pudiendo compensarse dando un error medio casi nulo que enmascara esos errores parciales.	$\overline{p-o}$	$(-\infty,+\infty)$	0
Error absoluto medio. Promedio entre las diferencias absolutas entre predicción y observación. Evita los problemas de compensaciones.	$\overline{\lvert p-o\rvert}$	$[0,+\infty)$	0
Raíz del error cuadrático medio. Raíz del promedio de los cuadrados de las diferencias entre predicción y observación. Evita los problemas de compensaciones, pero penaliza mucho errores grandes. Sus unidades coinciden con las de la magnitud.	$\sqrt{\overline{(p-o)^2}}$	$[0,+\infty)$	0
Coeficiente de correlación de anomalías. Correspondencia entre anomalías respecto de la climatología, en forma similar al coeficiente de correlación.	$r_{p-c,o-c}$	$[-1,+1]$	1

Tabla 15.3: Resumen de los scores básicos para variables continuas (ver texto).

15.2.5 Representaciones gráficas

Los scores o medidas se representan de muchos modos, y presentamos aquí tres de los más comunes, ilustrados en las Figuras 15.5 en la página siguiente, 15.7 en la página 215 y 15.6 en la página siguiente.

Distribución espacial (Figura 15.5 en la página siguiente): un mapa con los valores del score, a menudo con un código de colores; los puntos irán coloreados según el valor del score en cada estación donde hubo observación contrastable con el modelo. Contesta a las preguntas del tipo *¿en qué regiones el modelo tuvo error positivo, es decir, sobrestimó el valor observado?*, *¿es homogéneo el comportamiento del modelo?*, *¿hubo una densidad razonable de estaciones para contrastar?*.

Evolución con el alcance de la predicción (Figura 15.6 en la página siguiente): se despliega la variación del valor del score con los alcances de la predicción. El modelo, supuestamente, va perdiendo calidad según avanza el alcance, y por tanto, el MSE debe crecer, etc. Igual que con las series temporales, el valor del score variando con el alcance de la predicción puede ser un atributo de calidad del modelo en un lugar o promediado en un dominio. Así mismo, pueden mostrarse scores de diferentes modelos, alcances, pasadas, lugares, etc, en diferentes colores o estilos, para comparación.

Serie temporal (Figura 15.7 en la página 215): se despliega la variación del valor del score con los días, meses, años o décadas. El valor del score puede ser un atributo de calidad del modelo en un lugar o promediado en un dominio. Pueden mostrarse scores de diferentes modelos, alcances, pasadas, lugares, etc, en diferentes colores o estilos, para comparación.

Serie temporal con media móvil. A menudo puede hacerse una media móvil temporal para filtrar ruido de corto plazo. En la Figura 15.8 en la página 215, la curva verde representa la calidad de ciertas predicciones del modelo ECHRES con variación diaria. La variabilidad de alta frecuencia impide comprobar tendencias de más largo plazo, es una curva *ruidosa*. Para ello, puede usarse un filtro estadístico llamado *media móvil* (*moving average* o *running mean*, en inglés), consistente en sustituir el dato de cada día por una media de los datos en una ventana de cierto tamaño centrada en ese día, de modo que se consiguen suavizar los rangos de alta frecuencia de la señal, preservando la variabilidad de frecuencia menor. El tamaño de la ventana de media móvil se ajusta para resaltar las frecuencias de mayor interés. En la Figura 15.8 en la página 215 la curva azul muestra una media móvil semanal, mientras que la curva negra muestra una media móvil mensual. Pueden observarse los distintos grados de filtro de frecuencia y cómo las tendencias de largo plazo quedan mucho más claras.

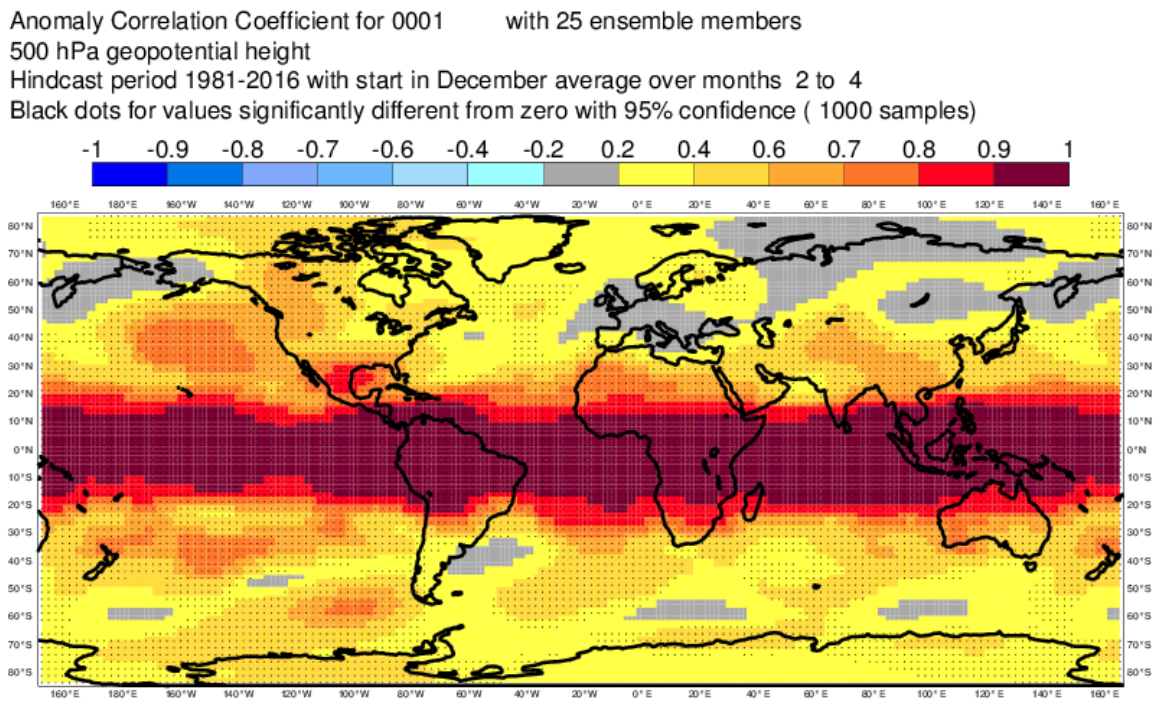

Figura 15.5: Representaciones gráficas habituales de los scores estadísticos que muestran información sobre diferentes atributos de calidad de los modelos. Aquí, un mapamundi con la distribución del coeficiente de correlación de anomalías (ACC de sus siglas en inglés). ECMWF.

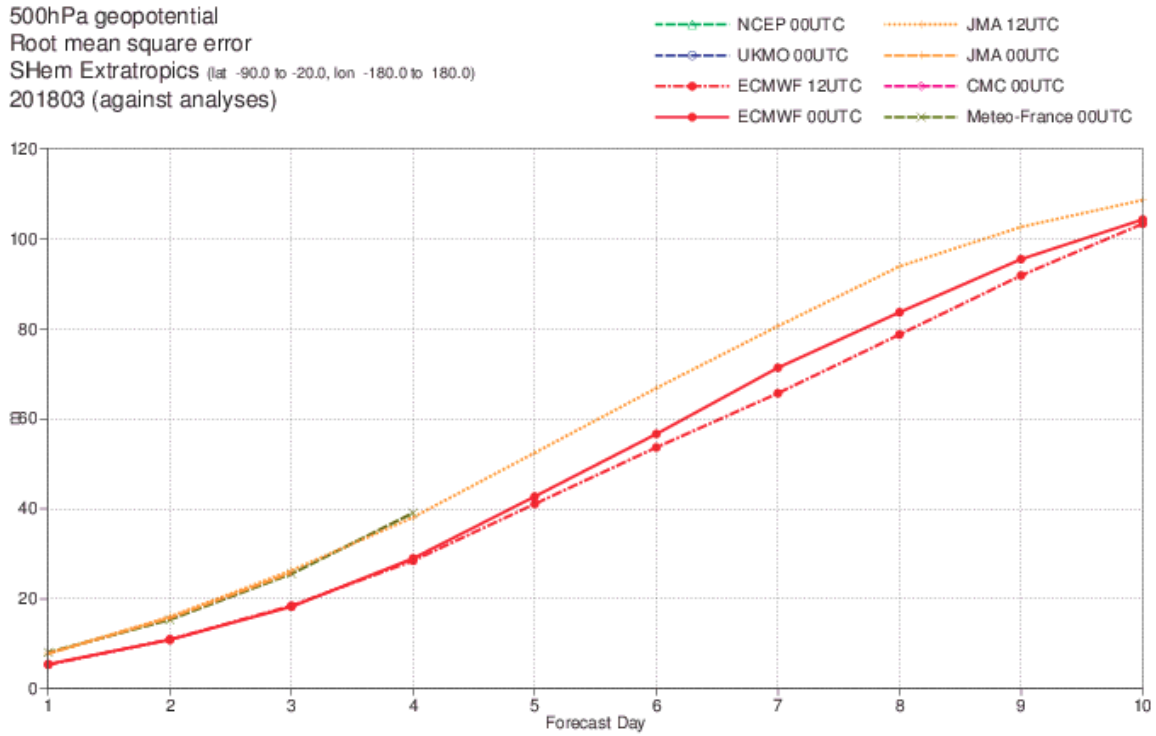

Figura 15.6: Evolución de la calidad con el alcance de la predicción, en términos de error cuadrático medio de predicciones de altura geopotencial en 500 hPa de varios modelos, en la franja extratropical sur, para marzo de 2018. El error aumenta con el alcance predictivo de forma natural. ECMWF.

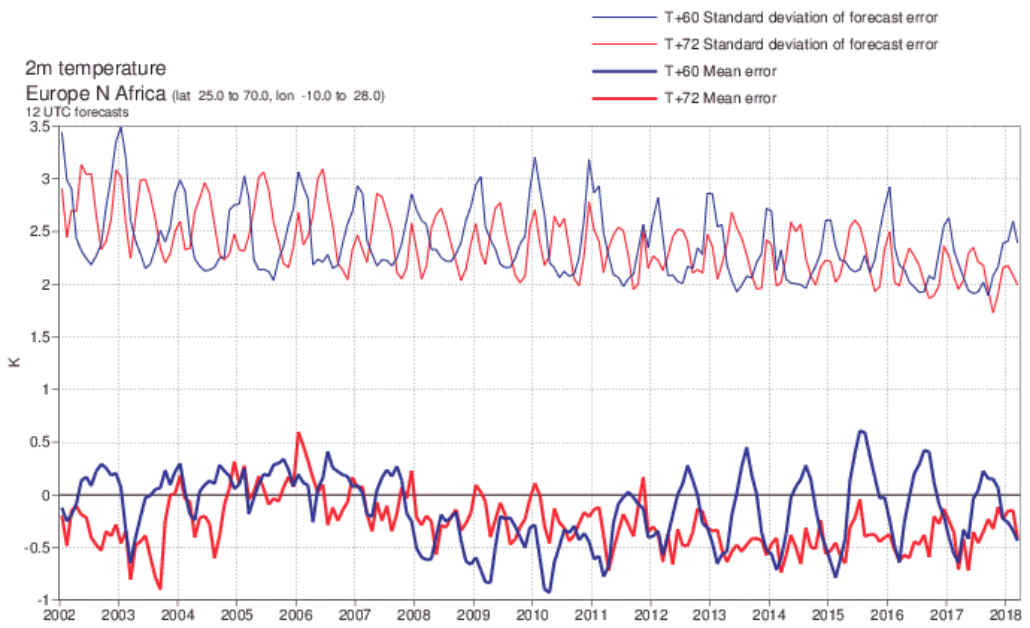

Figura 15.7: Serie temporal que abarca desde 2002 hasta el presente, mostrando el error medio (abajo, signo oscilante) y una variante del error cuadrático medio (arriba, siempre positiva) de las predicciones de temperatura del ECHRES. En azul, el alcance T+60 y en rojo el alcance T+72 h. Puede observarse que el error cuadrático medio va disminuyendo con los años. La comparación de calidad entre los dos alcances predictivos requiere analizar la hora del día y el comportamiento de la variable meteorológica en cuestión (ECMWF).

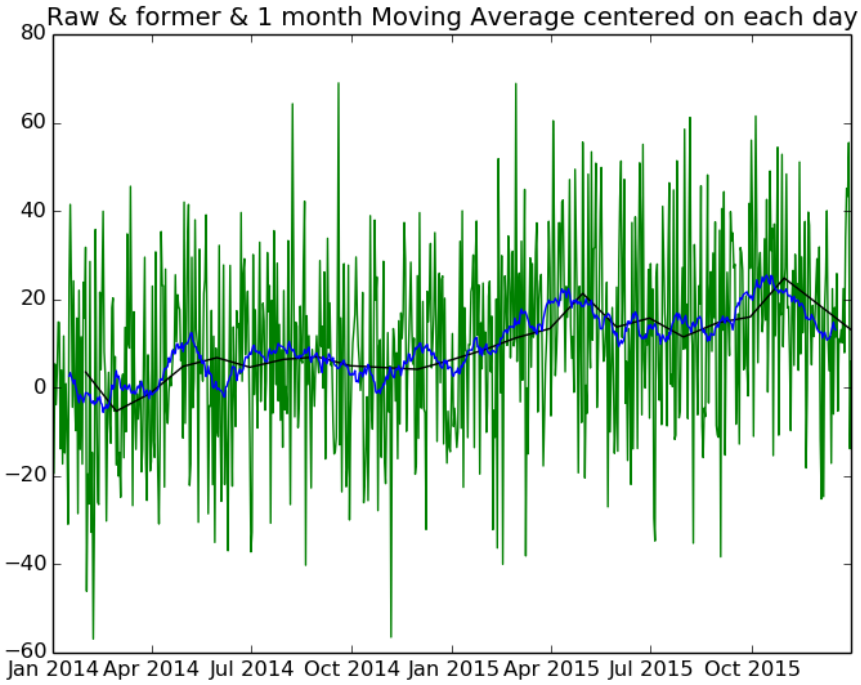

Figura 15.8: Serie temporal del valor diario de una cierta métrica (verde) que abarca dos años, 2014 y 2015. El ruido impide ver la tendencia, pero en las medias móviles de una semana (azul) y de un mes (negro) se filtra la alta frecuencia y emergen las tendencias de baja frecuencia (ver texto).

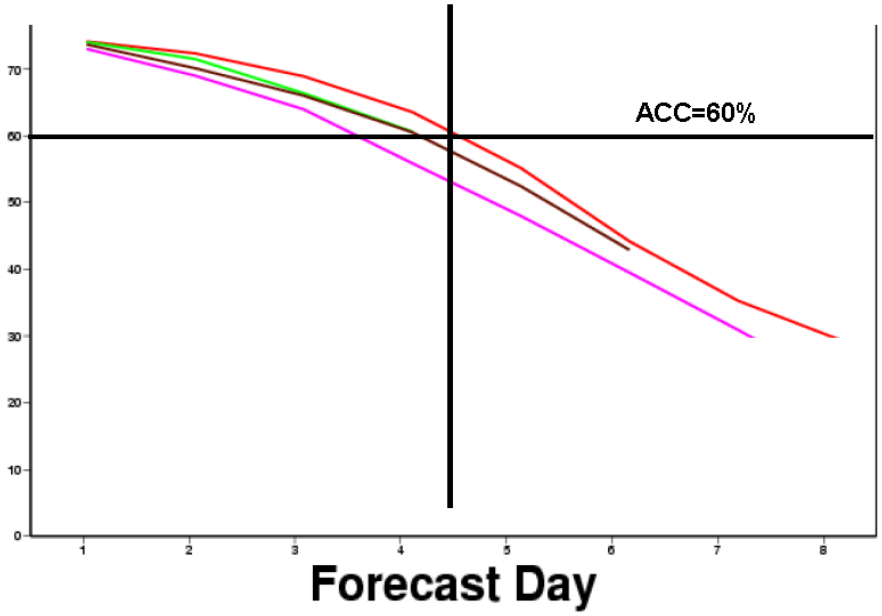

Figura 15.9: Evolución del coeficiente de correlación de anomalías (ACC) medio con el alcance predictivo (Forecast Day) que, seleccionando un umbral de calidad mínimo, define un límite de predecibilidad: para el modelo rojo sería, aproximadamente, cuatro días y medio. ECMWF.

15.3 El límite de predecibilidad

El límite de predecibilidad es un concepto clave en verificación de la predicción y, en general, en el mundo de la meteorología. Por un lado, es un concepto muy intuitivo por el que, de hecho, la sociedad suele formular preguntas del tipo ¿hasta qué día son fiables las predicciones? ¿he oído que hasta el quinto, verdad? Por otro lado y, dentro del marco más estrictamente teórico de la verificación, el concepto o atributo puede medirse de forma muy precisa utilizando una medida de calidad adecuada y viendo su evolución con el alcance predictivo. De este modo, puede tenderse un puente entre la verificación puramente objetiva y cuantificable y la idea social de la validez en el tiempo de las predicciones: una conexión entre el lenguaje científico o técnico y el social.

En la Figura 15.9 se muestra la evolución con el alcance predictivo (Forecast Day) del coeficiente de correlación de anomalías (ACC) de cuatro modelos distintos correspondientes a curvas de diferentes colores. El ACC es una medida orientada positivamente, es decir, cuanto más ACC, mejor correlación tiene la anomalía del modelo con la de la observación y, por ende, mejor calidad presenta el modelo. Como es natural, el ACC tendrá tendencia a decaer con el alcance predictivo: cuanto más lejos el horizonte de predicción, peor suele ser la misma. Esta tendencia es clara en los cuatro modelos mostrados en la Figura 15.9. Ahora elegimos, en principio arbitrariamente, un umbral de calidad según el ACC, por ejemplo 60 % y trazamos una recta horizontal en la gráfica precisamente en el 60 %. Esa recta define ahora dos regiones: pro encima, los modelos «son buenos» y, por debajo, «son malos». Pero, además, define una especie de *límite de predecibilidad*. Puede observarse que, para las diferentes curvas, ese umbral se traspasa en alcances predictivos distintos: el modelo rosa traspasa el 60 % en el D+3.7 aproximadamente, el marrón en el D+4.2 más o menos y el modelo rojo en el D+4.6. En este sentido, el modelo rojo permanece más tiempo con un ACC superior al 60 %. Podemos llamar a esos alcances umbrales *límites de predecibilidad* correspondientes a un ACC de un 60 %. Es decir, el límite de predecibilidad del modelo rojo será de 4 días y medio, más o menos y, podremos contestar a un periodista: «el modelo da predicciones de calidad hasta cuatro días y medio vista». Está claro que es decisivo elegir un ACC adecuado. Hasta 2008, en el ámbito europeo se venía trabajando con un ACC de 60 % pero, dado que los modelos han mejorado enormemente su calidad y que las comparativas se tornaban difíciles, se decidió en el Comité Técnico Asesor de Expertos en Verificación del ECMWF (del que el autor formaba parte) cambiar ese umbral a 80 %.

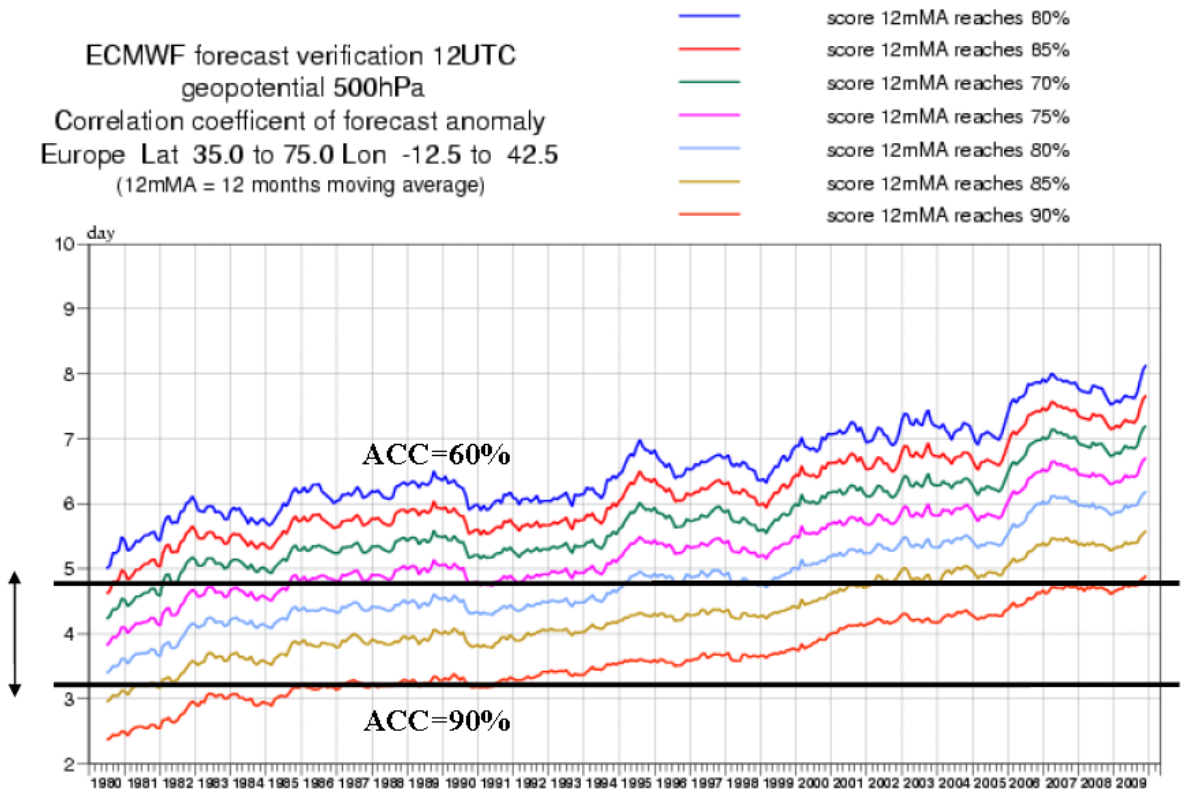

Figura 15.10: Límite de predecibilidad: serie temporal de límites de predecibilidad (ver texto). ECMWF.

Una vez que quedan bien definidos los límites de predecibilidad de los distintos modelos para distintos dominios, variables meteorológicas, épocas del año, etc., entonces pueden construirse series temporales de límite de predecibilidad, como la mostrada en la Figura 15.10. Está basada en el modelo determinista del ECMWF, el ECHRES (sec. 19.2 en la página 291). El eje vertical corresponde al límite de predecibilidad: cuanto más alto, mejor. El eje horizontal son los años de la serie temporal. Las curvas muestran la evolución de las predicciones de altura geopotencial en 500 hPa para la pasada de 12 UTC, en un dominio europeo (35-75º N 12.5º W-42.5º E). Cada punto corresponde a una media móvil de 12 meses centrada en el mes, lo que elimina la señal de alta frecuencia (variabilidad mensual), muy ruidosa y muestra mejor las tendencias a largo plazo (variabilidades mensuales suaves y, sobre todo, variabilidades estacional y anual). Cada color corresponde a un umbral dado de ACC: azul 60 %, rojo 65 %, verde 70 %, rosa 75 %, cían 80 %, amarillo 85 % y naranja 90 %. Pueden destacarse varias pautas:

- El límite de predecibilidad va aumentando con los años: los modelos van mejorando.
- Con el umbral de ACC=80 % podemos ver que la predecibilidad en 1980 tenía un límite de unos 3.5 días, en 1990 4.5 días, en 2000 5 días aproximadamente y en 2009 algo más de 6 días. Se suele decir que *en cada década se ha ganado de un día a día y medio de predecibilidad*, según la variable meteorológica estudiada. Las rectas negras horizontales muestran cómo, abajo, la predecibilidad para un ACC de 85 % en 1981 (algo más de 3 días), se alcanza en 1986 para un ACC del 90 %. La recta negra de más arriba muestra que la predecibilidad en 1980 para ACC=65 % era de 4.6 días aproximadamente y, esa predecibilidad, se alcanza en 2007 para ACC=90 %.
- La tendencia tiene altibajos: hay años donde aumenta relativamente menos o, incluso, disminuye. Esto se debe, en general, a la variabilidad natural de la predecibilidad: hay años más predecibles que otros. Hoy en día se puede relacionar esa predecibilidad con patrones de variabilidad conocidos (ver por ejemplo sec. 29.2.1 en la página 484 y sec. 29.2.2 en la página 485) o con teleconexiones.
- A más alto el umbral dado de ACC, más bajo es el límite de predecibilidad. Obsérvese, por ejemplo, que para ACC=90 % el límite de predecibilidad en 2009 termina en casi 5 días, mientras que para ACC=60 % termina en algo más de 8 días.

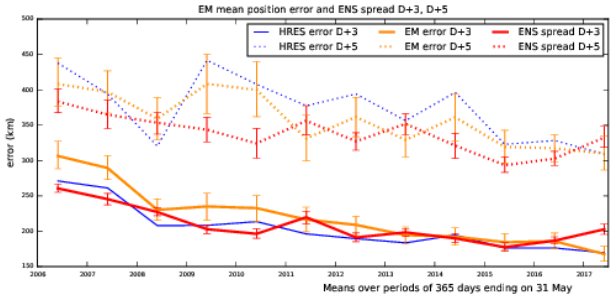

Figura 15.11: Series temporales del error de ECHRES (azul), el error del promedio del ECENS (amarillo) y de la dispersión de ECENS (rojo), en D+3 (trazo continuo) y D+5 (trazo punteado). Las barras de error despliegan la incertidumbre muestral con una cierta significación estadística. En los casos en que las barras superan la diferencia vertical entre dos curvas, no se puede afirmar con significación que un modelo muestra mejor o peor calidad que el otro. ECMWF.

15.4 Visión crítica y perspectivas

La mejora en recursos computacionales, la programación orientada a objetos y las mejoras en la aplicación de bases de datos han dado un buen empujón a la verificación de modelos atmosféricos. En las últimas décadas la comunidad científica dedicada a la verificación ha podido abordar importantes aspectos que tienen un fuerte impacto en la interpretación de las medidas de verificación, así como en la introducción de nuevas concepciones.

Compromiso entre agrupación y estratificación. Para poder interpretar la verificación las muestras deben tener significación estadística. Se busca así un compromiso entre la agrupación en muestras suficientemente grandes, pero a la vez estratificando el conjunto total en muestras que deben ser grandes pero climatológicamente homogéneas. E. g. en muestras heterogéneas puede haber compensación en el error medio, lo que puede conducir a información engañosa o confusa. Tradicionalmente se hace una estratificación estacional de los datos, lo que da cuenta de la variabilidad estacional, pero no de la variabilidad diaria [26, 27, 28, 48].

Verificación dependiente del régimen atmosférico. En el caso ideal la estratificación de los datos puede intentar hacerse por distintos regímenes atmosféricos, usando los actuales métodos de agrupamiento, mejorando así la tradicional estratificación estacional [13, 49].

Incertidumbre muestral. Las parejas {predicción, observación} disponibles son, en la práctica, muestras relativamente pequeñas y, desde el punto de vista estadístico, las medidas o scores computados son solamente estimadores muestrales de los valores poblacionales, y llevan una incertidumbre relacionada con este proceso de estimación. La información de verificación debe ir acompañada de esta incertidumbre muestral, en forma de barras de error o intervalos de confianza [3, 4, 12, 18]. Ver Figura 15.11.

Escalas espaciales de modelos y observaciones. Este es un punto clave que se describe con pleno detalle más adelante en este capítulo (sec. 15.5 en la página siguiente). Como ejemplo, la doble penalización es un problema bien conocido por los modelizadores. El desarrollo relativamente reciente de los llamados métodos espaciales de verificación [1, 19]que dan cuenta de los patrones geométricos (e. g. CRA[11], MODE[10], SAL[46]) están mostrando excelentes resultados. Conducen a un marco de verificación diagnóstica, cercana a la verificación subjetiva en el sentido de que brinda información que puede ayudar a los modelizadores y predictores a identificar errores en el modelo, desde el punto de vista meteorológico. Cercana también a la verificación objetiva en el sentido de que se basa en resultados cuantitativos.

Error en las observaciones. Se asume tradicionalmente que el error observacional es despreciable comparado con el error predictivo. Si bien es generalmente cierto para plazos de predicción medio-largos, no lo es para el corto plazo. Si se tiene en cuenta este error en forma de incertidumbre los resultados pueden ser sorprendentes [8, 42].

Fenómenos extremos y fenómenos adversos. Fenómenos extremos (*extreme weather*) son eventos raros, con frecuencia climatológica pequeña y la verificación de los mismos está todavía en una etapa de poca madurez. Los fenómenos adversos (*severe weather*) son aquellos con impacto social y económico, y su verificación debe incluir información no meteorológica, tal como coste de medidas preventivas si se hace caso de la predicción o pérdidas por daños, etc. [14, 17, 30, 31, 40, 44, 51]

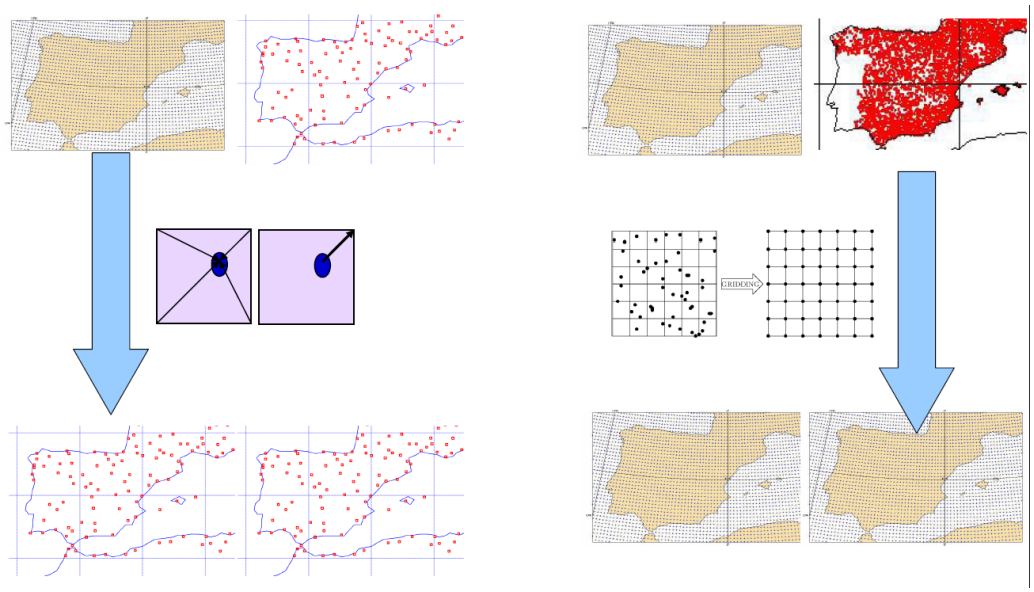

Figura 15.12: Comparación de dos métodos habituales para adaptar representaciones espaciales. A la izquierda: interpolación del modelo a los puntos de observación. Derecha: mallado *(gridding o upscaling) de las observaciones a la malla del modelo (ver texto).*

15.5 Métodos espaciales

Comparar observaciones y predicciones es como comparar peras con manzanas: a menudo son representaciones de la atmósfera de naturalezas diferentes y necesitaremos transformar al menos una de ellas. Esta transformación requiere procesos como la interpolación e implica problemas de representatividad, correlación, ruido, etc.

15.5.1 Predicciones de precipitación

Un ejemplo típico para ilustrar estas dificultades es la verificación de predicciones de precipitación de un modelo determinista (en inglés *Quantitative Precipitation Forecast*, QPF). Las redes climáticas europeas de alta densidad proveen datos observados de precipitación acumulada en 24 horas. Estas estaciones están distribuidas en el espacio de forma irregular. Sin embargo, las predicciones del modelo están regularmente distribuidas en el espacio (*malla* o *grid*). La precipitación observada es puntual y sin embargo la precipitación prevista es *areal*. Si la red de estaciones no cubre con densidad suficiente la malla del modelo (Figura 15.12, izquierda), entonces pueden interpolarse los valores del modelo a los puntos de observación y trabajar las comparaciones en esos puntos[41]. En este caso debe guardarse especial cuidado con el impacto de la posible falta de consistencia estadística debido a la correlación espacial entre estaciones próximas. Por otro lado, si la densidad de estaciones es suficiente (Figura 15.12, derecha), puede calcularse una estimación de la precipitación observada (en inglés *Quantitative Precipitation Estimate*, QPE) usando una técnica de aumento de escala (upscaling[16]) para *mallar* las observaciones en la *malla* del modelo. En cada celda se toman las observaciones que caigan dentro y se calcula una cantidad representativa (e. g. un promedio pesado) de carácter areal. El método elegido (interpolación o upscaling) depende en general de la densidad de observaciones (ver Figura 15.12). PAIMAZUMDER y MÖLDERS (2009[37]) estudiaron el impacto de la densidad y el diseño de las redes de observación en promedios regionales usando datos reales sobre Rusia. Encontraron que, generalmente, las redes reales subestiman las medias regionales de presión al nivel del mar, velocidad del viento y precipitación, mientras que sobrestiman la temperatura a dos metros, la radiación entrante de onda corta y la temperatura del suelo.

En la comparación de rendimiento de la QPF de dos modelos diferentes pueden aparecer problemas adicionales si el tamaño de celda es distinto. Los dos modelos, de diferentes resoluciones, representan estructuras de precipitación cada uno en su escala. Si las observaciones se mallan a la resolución más fina, entonces el modelo más grueso estará penalizado. Si, por el contrario, las observaciones se mallan a la resolución más gruesa, entonces el modelo fino no tiene oportunidad de demostrar su rendimiento debido a la mayor resolución.

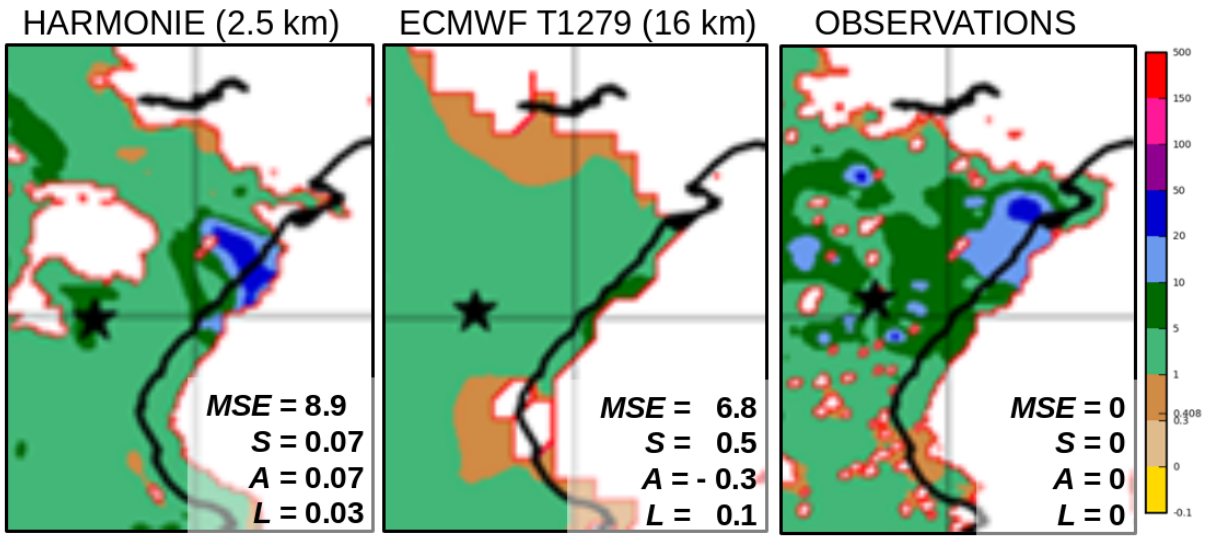

Figura 15.13: Ilustración del problema de la doble penalización *(*double penalty*) con un caso real (ver texto). Imagen de* ARANCHA AMO*, AEMET 2013.*

15.5.2 Doble penalización

Otro ejemplo típico del impacto importante de la escala de predicciones y observaciones es el famoso problema de la doble penalización, problema bien conocido por los modelizadores (Figura 15.13). Tomamos, por un lado, un modelo de alta resolución, e. g. HARMONIE-AROME, Figura 15.13 izquierda y, por otro lado, un modelo global de menos resolución, e. g. el modelo determinista del ECMWF, ECHRES (sec. 19.2 en la página 291), Figura 15.13 centro. Y comparamos ambos con las observaciones de precipitación, e. g. la red termopluviométrica de la Agencia Estatal de Meteorología (AEMET) (sec. 9.5 en la página 122), mallada mediante un proceso de upscaling, Figura 15.13 derecha. Computando el error cuadrático medio (MSE) de ambos modelos con respecto a la observación, resulta que el modelo global obtiene menos MSE y podría parecer, a primera vista, que disfruta de más calidad. Si evaluamos cualitativamente (información diagnóstica), vemos que HARMONIE-AROME resuelve bastante bien el patrón de precipitación, dando un núcleo importante de precipitación un poquito más al sur que el núcleo observado, mientras que ECHRES da información más suave en general. HARMONIE-AROME es penalizado doblemente, en aquellos puntos donde arriesgó dando lluvia que no se observó y, viceversa, también en aquellos puntos donde no daba lluvia pero llovió. De ahí que el error (MSE) de HARMONIE-AROME resulte mayor, porque sufre dobles penalizaciones. Pero, desde un punto de vista diagnóstico, sabemos que tiene gran valor. ¿Cómo solucionar este problema? Necesitamos métodos de verificación diagnósticos (cuantitativos pero que aporten información diagnóstica). Esos son los métodos llamados espaciales.

15.5.3 Métodos diagnósticos espaciales

Los métodos espaciales de verificación, de desarrollo relativamente reciente, tuvieron un importante eco internacional gracias al Proyecto de Intercomparación de Métodos Espaciales de Verificación [1, 19]. Estos métodos dan cuenta de los patrones espaciales y están mostrando prometedores resultados. Algunos ejemplos son CRA [11], MODE [10] y SAL [46]). Ofrecen un marco de verificación diagnóstica, cercana a la verificación subjetiva en el sentido de que brinda información que puede ayudar a los modelizadores y predictores a identificar errores en el modelo desde el punto de vista meteorológico y, cercana a la objetiva, en el sentido de que se basa en resultados cuantitativos. Resuelven el problema de la doble penalización y por tanto, permiten contrastar adecuada y justamente la calidad de los modelos de alta resolución. En el ejemplo ilustrado en la Figura 15.13 se muestran, además del MSE, resultados obtenidos con el método SAL [46], que arroja tres cantidades que describen, respectivamente, estructura (S), amplitud (A) y localización (L) del patrón de precipitación previsto en relación con el observado. Son medidas de error, es decir, cuanto más cercanas a cero mejor. Inspeccionando los valores comparativos entre HARMONIE-AROME y ECHRES, podemos inferir que HARMONIE-AROME describe los patrones de precipitación con más calidad que ECHRES.

15.6 ¿Son perfectos los SPC?

Si la verificación determinista, como hemos visto, lleva consigo toda una problemática inherente en su teoría y práctica, la verificación probabilista presenta dificultades superiores desde un principio. Los SPC trabajan con funciones de densidad de probabilidad (probability density functions, PDF, sec. 12.2.1 en la página 158 y sec. 13.6 en la página 176) y proporcionan predicciones en forma de probabilidades y sus derivados (cap. 27 en la página 401).

Retomando el ejemplo cotidiano de la sección 15.2.3 en la página 211, un modelo determinista predice 4º C en Sevilla y después se observan 3º C: el error ha sido de 1º C. Sabemos que no debemos extrapolar esa medida de verificación para generalizar sobre la calidad de el modelo, porque hay que usar muestras con significación estadística, i. e. grandes y homogéneas. Pero, al menos, podemos comparar algebraicamente la predicción y la observación: $4-3=1$. Pasando al mundo probabilista, una predicción para Sevilla sería La probabilidad de que caigan más de 5 mm es del 30 %. Al día siguiente se observan 3 mm. ¿Cómo comparamos la predicción con la observación? Si en los sistemas deterministas la significación estadística era importante y por ende se deben tomar muestras grandes, tomar muestras grandes en ensembles es imprescindible, pues carece de sentido comparar $Prob(>5)=0{,}3$ con $ob=3\,mm$). Surgen por tanto, numerosas preguntas, que podemos resumir en dos:

- ¿Qué es una buena PDF?
- ¿Qué se le pide a un buen SPC?

En términos generales, hay tres grandes aspectos: la calidad de los miembros individuales, la consistencia con las observaciones en el flujo a gran escala y la respuesta a los eventos binarios en los parámetros de tiempo sensible. Abordaremos estos tres aspectos, por su importancia, en tres secciones separadas.

15.7 Calidad de los miembros individuales

Un requisito fundamental en un SPC es que las predicciones de los distintos miembros sean equiprobables para poder así calcular probabilidades a partir de una PDF bien construida. Se trata de un requisito previo que se consigue en la fase de desarrollo de un ensemble. Cuando la calidad de los miembros es equiparable, entonces los pesos de los miembros son iguales para calcular probabilidades (sec. 13.7.2 en la página 191).

Por otro lado, el promedio del SPC, aunque hemos subrayado que no tiene por qué ser una situación meteorológica plausible (sec. 13.6.3 en la página 178), curiosamente muestra una calidad determinista estadísticamente superior al resto de miembros [32, 36, 47, 50].

Acerca de la calidad individual de los miembros, es costumbre verificar la evolución según el alcance de predicción de variables dinámicas. Por ejemplo, la Figura 15.14[15] muestra el error medio, mean error, ME (grupo inferior de curvas, signos diversos) y la raíz cuadrada del error cuadrático, root mean squeare error, RMSE (grupo superior de curvas, sólo positivo) en la presión (MSLP) para los miembros (trazo fino) y para el promedio (trazo grueso) del ensemble AEMET-SREPS (cap. 21 en la página 313). Nótese cómo el RMSE del promedio es inferior al del resto de miembros: una predicción que puede no ser plausible es, sin embargo, estadísticamente consistente.

En las épocas de adaptación a la novedad de los SPC (1995-2005) era habitual el debate sobre la comparación de calidad entre un modelo determinista y su SPC correspondiente (si tal correspondencia se daba). Por ejemplo, tomando el modelo determinista del ECMWF, el ECHRES (sec. 19.2 en la página 291) y el SPC correspondiente, el ECENS (sec. 19.3 en la página 293), el promedio del ECENS era mejor estimador que ECHRES más allá del D+3/D+4 para variables diversas como Z500, T2m o la precipitación[38, 39].

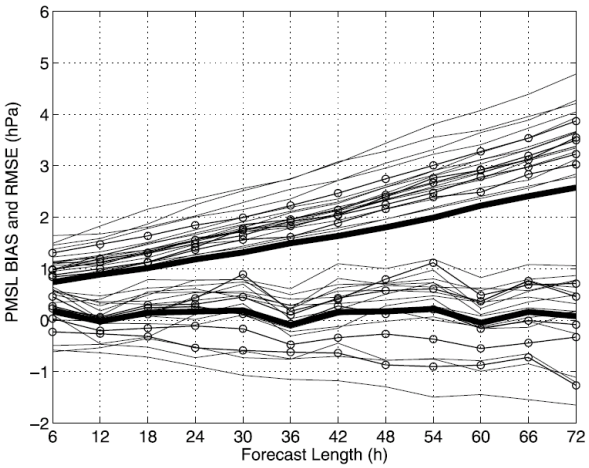

Figura 15.14: Evolución con el alcance de la predicción de los errores en predicciones de presión de los miembros de un ensemble, en trazo fino, así como de su promedio, en trazo grueso (ver texto). [15] TellusA, CC BY 4.0.

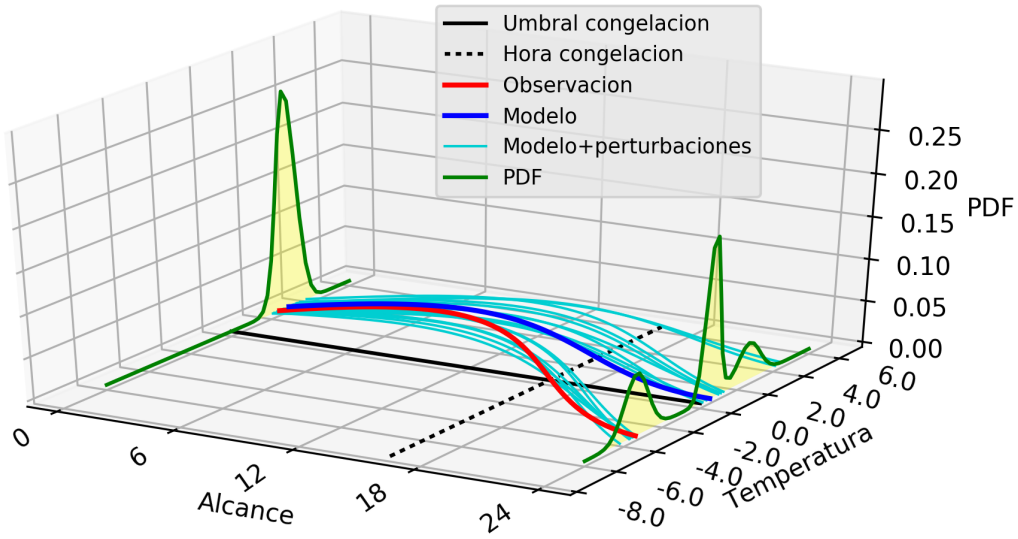

Figura 15.15: Evolución de la PDF observada hacia una PDF prevista (ver texto).

15.8 Consistencia del flujo a gran escala

En la representación del flujo atmosférico a gran escala y en un dominio suficientemente grande un SPC debe ser consistente con las observaciones. Usualmente se compara con un análisis, para garantizar una homogeneidad entre zonas con diferentes coberturas y se comparan parámetros dinámicos de capas medias-altas, e. g. altura geopotencial en 500 hPa, que tienen una variabilidad suave en las escalas de interés.

Recordando el ejemplo de la predecibilidad en la vida cotidiana, un ensemble consistente con las observaciones es aquel cuya PDF es capaz de capturar la realidad (ver Figura 15.15): que la observación (curva roja) caiga siempre dentro del abanico de probabilidades de la PDF prevista (zonas amarillas calculadas con las curvas azules). Artificialmente, podríamos engordar la PDF de modo que siempre capturara la realidad, e. g. podríamos decir la temperatura para mañana en Sevilla estará entre 10 y 40º C, pero esa predicción dejaría de ser útil. Por otro lado, una predicción demasiado ajustada, e. g. entre 15.3 y 15.5º C, puede parecer útil, pero es raro que capture la realidad. Se busca un compromiso entre una PDF de tamaño y forma razonable que capture la realidad y que sea útil.

Esta consistencia puede medirse de muchas formas. Presentamos aquí los histogramas de rango y los diagramas dispersión-error.

15.8.1 Histogramas de rango

Asumimos que en cada predicción el valor observado y los valores previstos por los diferentes miembros son *realizaciones independientes* del mismo proceso atmosférico y por tanto *equiprobables* (sec. 13.7.2 en la página 191). Esta asunción se cumple realmente solo cuando el histograma de rango es plano.

Rango de la observación. El *rango*, en sentido estadístico, de la observación es un número entero que denota la posición relativa que ocuparía ese valor en la lista ordenada de predicciones. Por ejemplo, un ensemble de cinco miembros predice en un momento y lugar los valores (18, 16, 14, 13, 15). Ordenando la lista resulta: (13, 14, 15, 16, 18). Si el valor de la observación es 14.5, ésta ocuparía el tercer lugar, que se denota como rango 2, dado que el primero es el rango 0: (13, 14, *14.5,* 15, 16, 18). Calculando el rango de la observación entre los N valores de predicción en todos los momentos y lugares, acumularemos esa información en el así llamado *histograma de rango* [2, 7, 21, 22, 23], usado para comprobar si el ensemble es estadísticamente consistente con las observaciones. Un histograma en forma de U indicará la presencia de *valores atípicos* (*outliers*) en los extremos, muestra de subdispersión, problema usual en los sistema(s) de predicción por conjuntos (SPC) operacionales. La forma de campana indicará sobredispersión, producto eventual de una varianza engordada en demasía por posproceso estadístico. Los histogramas inclinados muestran un sesgo predictivo: pendiente positiva (negativa) para un sesgo negativo (positivo). Un histograma plano indicará consistencia con las observaciones (Figura 15.16 en la página siguiente).

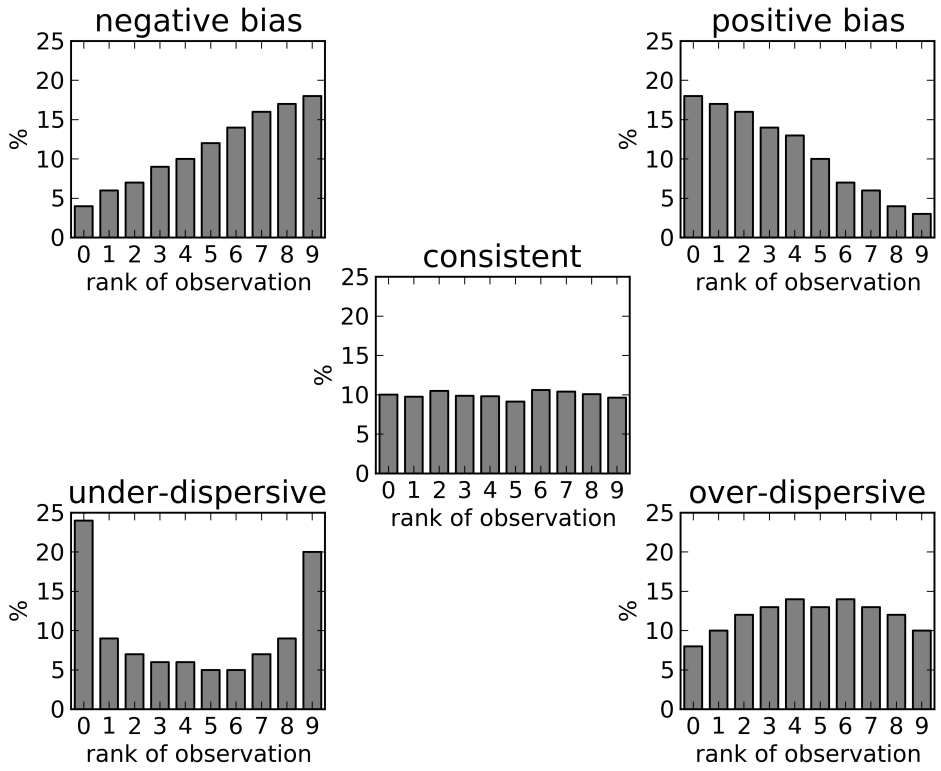

Figura 15.16: Histogramas de rango [6].

15.8.2 Diagramas dispersión-error

Por un lado, la dispersión del ensemble se mide usualmente con la desviación estándar con respecto al promedio. Por otro lado el error del ensemble se suele medir con la RECM del promedio, o del miembro de control si lo hay, con respecto al análisis. En un ensemble consistente con las observaciones esperamos que estas dos magnitudes, dispersión y error, crezcan con el alcance de predicción mostrando una relación aproximadamente lineal y de pendiente unidad, es decir, que crezcan a ritmos similares [5, 47]. Un buen ensemble muestrea las incertidumbres de los modelos (dispersión, PDF del ensemble) y cuantifica la predecibilidad atmosférica (error, PDF atmosférica) y es consistente si ambas cantidades crecen a ritmos similares. En ese caso se dice que está bien calibrado. En caso contrario, bien por subdispersión o por sobredispersión, el ensemble necesita calibrarse. La Figura 15.17 muestra un diagrama dispersión-error real. En el eje horizontal, el error del promedio del SPC con respecto al análisis. En el eje vertical, la desviación estándar de los miembros del SPC con respecto al promedio. La curva gruesa representa los valores dispersión-error para predicciones de mean sea level pressure -presión reducida al nivel medio del mar- (MSLP) del SPC AEMET-SREPS (cap. 21 en la página 313) durante un periodo dado en el dominio de Península y Baleares. Las curvas finas con símbolos corresponden a diversos subensembles de AEMET-SREPS. Aquellas curvas por encima de la diagonal indican sobredispersión y, por el contrario, por debajo de la diagonal indican subdispersión. Puede comprobarse que el SPC completo muestra, en general, mejor relación dispersión-error que los diversos subensembles.

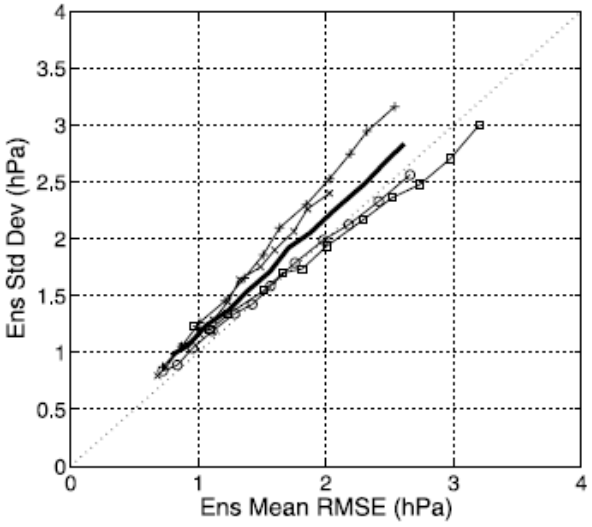

Figura 15.17: Diagrama dispersión-error. [15] TellusA, CC BY 4.0.

15.9 Eventos binarios

Para dar medidas de calidad de un SPC en parámetros de tiempo sensible, e. g. precipitación, temperatura, viento, que tienen mayor variabilidad en el espacio y en el tiempo que los parámetros dinámicos como la altura geopotencial o la presión, es mejor usar observaciones que usar análisis. Dos marcos de trabajo usuales son el marco de BRIER y el marco de la distribución conjunta [26, 48] basados ambos en la respuesta del SPC frente a *eventos binarios*, aquellos eventos que ocurren o no ocurren. Un evento binario siempre puede definirse para un parámetro continuo como la superación de un umbral, e. g. puede considerarse la superación del umbral de lluvia 1 mm para definir el evento *llueve o no llueve*. La calidad del SPC se describe en el espacio de las probabilidades con propiedades como fiabilidad, resolución, agudeza y discriminación, mientras que los beneficios de usar el SPC pueden medirse con el llamado valor (económico) relativo, que depende además de la relación coste/pérdida del usuario.

Climatología muestral. Para figurar adecuadamente los conceptos de esta sección imaginaremos un ejercicio de verificación de un hipotético SPC en un periodo de tiempo, e. g. tres meses, sobre una región, e. g. Baleares, evaluando el evento binario *temperatura por debajo de 0 ºC o por encima de 0 ºC* para cada día. Este evento se predice para una serie de puntos geográficos para cada día y, a toro pasado, se observa si la temperatura estuvo por encima, o no, de 0 ºC. Si se dispone e. g. de 40 estaciones meteorológicas, conseguiremos 40 datos para cada día, lo que en tres meses hace un total esperado de $90 \times 40 = 3\,600$ casos. Como en la realidad no se obtienen datos en todos los casos (fallos técnicos, etc.), la muestra real tiene 3455 datos. En cada caso el SPC emite una probabilidad de temperatura positiva, e. g. 10 %, etc. y tendremos la observación correspondiente, i. e. temperatura positiva o negativa. Este conjunto de casos forma lo que llamaremos *muestra* o *climatología muestral* de la verificación.

15.9.1 Marco general de Brier

En el llamado marco de BRIER se generaliza el concepto de error cuadrático medio (mean square error, MSE, sec. 15.2.3 en la página 211) extendiéndolo al espacio de las probabilidades. En el contexto de la predicción determinista, para cada caso de interés p y o son la predicción y la observación, e. g. 3 ºC de temperatura previstos y 2 ºC observados. Ahora, en el contexto de la predicción probabilista con un evento binario definido (e. g. temperatura superior a 0 ºC), p es la probabilidad prevista, e. g. 30 % ó 0.3 de que $T > 0$ºC (donde, en el caso de un SPC, el 30 % viene dado por la proporción de miembros que predicen $T > 0$ºC) y o es la observación o no del evento binario, definida como 0 ó 1 si el evento no ocurre o sí ocurre, respectivamente (e. g. si se observan 2 ºC entonces $o = 1$, si se observan -5 ºC entonces $o = 0$).

BRIER score (*BS*). Generalizando el error cuadrático medio (MSE) a este *espacio de probabilidades* se define el así llamado *índice de* BRIER o, en inglés, BRIER *Score* (*BS*):

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n)^2 \Longrightarrow BS = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n)^2 \tag{15.7}$$

donde N es el número de casos y, para cada caso n, p_n y o_n son la predicción y la observación en el marco de la predicción determinista y error cuadrático medio, mientras que p_n y o_n son la probabilidad prevista y la observación o no del evento en el marco de la predicción probabilista e índice de BRIER. Al ser un error, cuanto más pequeño el *BS*, mejores son las predicciones, se dice así que *BS* es un índice orientado negativamente. $BS = 0$ para un sistema de predicción perfecto y $BS = 1$ para un sistema de predicción «peor imposible» (es tan difícil hacerlo perfectamente mal como perfectamente bien).

Descomposición del BRIER score (*BS*). Se suele hacer una partición en el espacio de las probabilidades, e. g. intervalos de tamaño 0.1 y, en esa partición, el índice de BRIER puede descomponerse del modo siguiente:

$$BS = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}(p_n - o_n)^2 =$$

$$\tfrac{1}{N}\textstyle\sum_{i=1}^{I} n_i (p_i - o_i)^2 - \tfrac{1}{N}\sum_{i=1}^{I} n_i (s - o_i)^2 + s(1-s) \tag{15.8}$$

donde I es el número de intervalos de la partición, p_i y o_i tienen el mismo significado que anteriormente y s es la frecuencia de ocurrencia del evento o frecuencia de observación del evento. Las diferentes componentes tienen un significado e interpretación que abordaremos más adelante en esta sección. La así llamada *distribución conjunta* [33] de probabilidades previstas p_i y observaciones correspondientes o_i queda definida por las distribuciones correspondientes en el espacio de las probabilidades, que permiten trabajar en un marco orientado a distribuciones [26, 28].

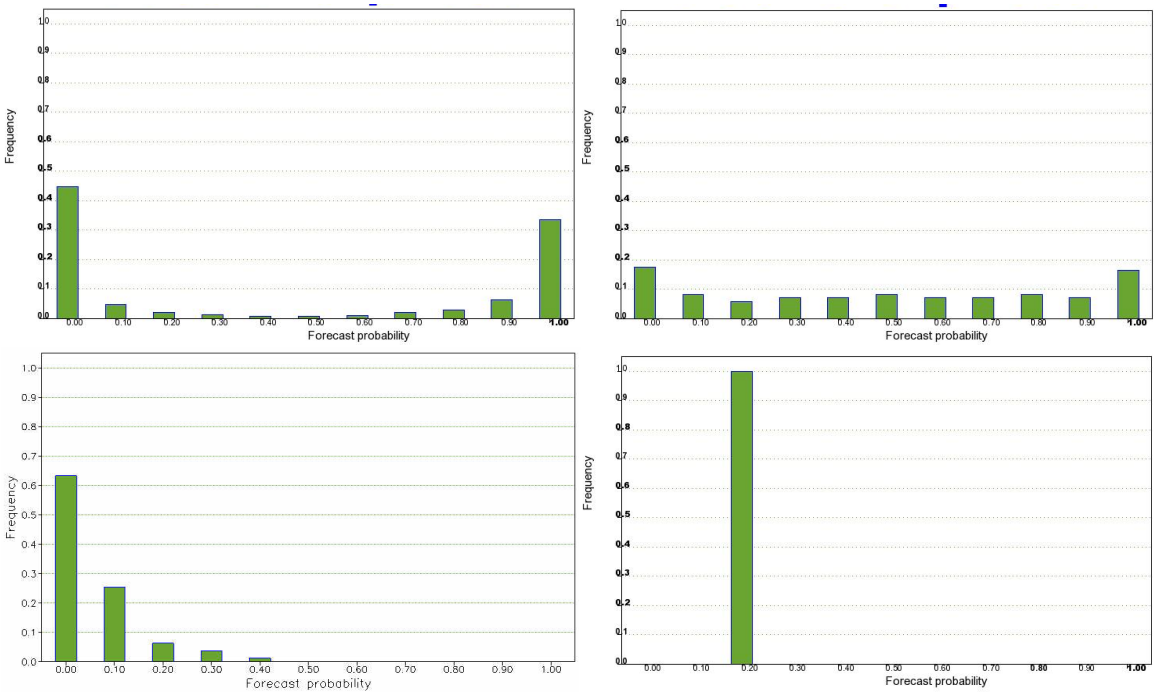

Figura 15.18: Agudezas de diferentes sistemas predictivos. Preferiremos un sistema con histograma de agudeza en forma de U *(arriba izquierda) donde se emiten más 0, 10, 90 y 100 %. El sistema de arriba a la derecha muestra, sin embargo, grandes cantidades de otros intervalos como el 40 o el 50 %, indicando una agudeza pobre. Tratándose de un evento relativamente raro (abajo izquierda), será normal emitir pocos 90 y 100 % y, en general, mostrar el histograma de agudeza con una clara pendiente negativa; no es que el sistema tenga poca agudeza, sino que el fenómeno es raro (e. g. superación de un umbral alto de viento). La climatología muestral (abajo derecha), como sistema predictivo teórico, tiene muy poca agudeza: emite sólo una probabilidad (la de la frecuencia de ocurrencia) por lo que su histograma tiene una única barra.*

15.9.2 Frecuencia de ocurrencia

En la muestra seleccionada de 3 455 casos el evento *temperatura positiva* tuvo lugar en una cierta parte de ellos, en 864 de los casos. La *frecuencia climatológica de ocurrencia observada de la climatología muestral* es el valor $\frac{864}{3455} = 0{,}25 = 25\,\%$ o, expresado de otro modo, temperatura positiva en el 25 % de los casos. Este valor es importante pues sirve como referencia para las propiedades que se definen a continuación.

La frecuencia de ocurrencia como predicción probabilista figurada. La predicción probabilista más sencilla posible, si se pudiera atisbar en el futuro para averiguar la frecuencia de ocurrencia en el periodo de interés, consistiría simplemente en emitir, para todas las localidades y todos los días, la misma probabilidad de 25 %. Estadísticamente hablando tiene sentido, dado que en el 25 % de los casos la temperatura fue positiva. Como predicción teórica, la *frecuencia de ocurrencia* tiene un sentido constructivo conceptual que veremos en los apartados siguientes.

15.9.3 Agudeza

Sin atender al valor de las observaciones que correspondieron a las predicciones, es decir, mirando solo las predicciones, la *agudeza* de un sistema predictivo probabilista tiene que ver con *cuánto se arriesga* el sistema desde el punto de vista del usuario. Si pensamos en el evento binario *temperatura positiva o negativa* el sistema da, en distintos lugares del dominio y momentos del periodo, distintas probabilidades de temperatura positiva: 100 %, 90 %, etc. Como usuarios, preferimos muchos 0, 10, 90 y 100 junto con pocos 40, 50 y 60. Saber que se espera un 50 % de probabilidad de temperatura positiva me ayuda poco y, sin embargo, un 90 ó un 100 % me ayudan mucho más. La agudeza puede medirse mediante un histograma de frecuencias de probabilidades: para cada intervalo de probabilidad, una barra vertical con las veces que se emitió esa probabilidad: cuántas veces 0 %, cuántas 10 %, etc. De un vistazo sabremos si el sistema emite más veces 0 y 100 % o más veces 50 %. En la Figura 15.18 se ilustran varios ejemplos.

q (%)	p	qp (perfecto)	o (real)	f (%)
0	2027	0	17	1
5	143	7	12	8
10	104	10	12	12
15	65	10	17	26
20	55	11	6	11
25	51	13	13	25
30	45	14	7	16
35	40	14	13	32
40	38	15	22	58
45	40	18	13	32
50	35	18	19	54
55	39	21	24	62
60	35	21	21	60
65	24	16	18	75
70	32	22	21	66
75	36	27	28	78
80	50	40	46	92
85	48	41	34	71
90	58	52	54	93
95	95	90	90	95
100	395	395	385	97

Tabla 15.4: Tabla con probabilidades previstas y frecuencias de observación condicionadas correspondientes en una muestra de verificación real (temperaturas previstas por AEMET-γSREPS y observadas en estaciones surface synoptic observations (SYNOP) en 3 455 casos), donde q *son los intervalos de probabilidad,* p *los casos previstos en cada intervalo,* qp *los casos en cada intervalo donde se esperaría observar el evento si el sistema fuera perfectamente fiable,* o *el número de casos donde realmente se observa el evento en cada intervalo y, finalmente,* f *es la frecuencia relativa de esos casos. El contraste entre* q *y* f *nos indica la fiabilidad del sistema.*

15.9.4 Fiabilidad

Si nuestro sistema tiene 20 miembros, podemos pensar en las probabilidades emitidas separadas en intervalos de 5 en 5 %: 0, 5, 10, ... 95, 100 % (a menudo se simplifica haciendo la partición en 10 intervalos). Para cada intervalo tendremos una serie de casos en los que habremos emitido esa probabilidad y sólo en un subconjunto de ellos realmente habrá habido temperatura positiva. La pregunta *¿en cuántos habrá habido temperatura positiva dentro de cada intervalo?* conduce a una respuesta de carácter estadístico, que depende del intervalo en sí. Si, según los datos de la Tabla 15.4, de los 3 455 casos, se han emitido 395 predicciones con un 100 %, esperamos estadísticamente que la temperatura haya sido positiva realmente en 395 de los mismos, ya que estábamos emitiendo una probabilidad de 100 %. Si se han emitido 58 casos con un 90 %, esperamos estadísticamente que haya llovido realmente en un 90 % de los mismos, es decir, en $58 \times 0{,}9 = 52$ casos. Y así sucesivamente hasta llegar al 0 %. Tendríamos así los casos esperados. Ahora se observa si la temperatura fue positiva realmente y, *a posteriori*, pueden cotejarse los casos esperados con los observados. En un sistema predictivo perfectamente *fiable*, los esperados coinciden con los observados, e. g. hemos emitido probabilidad del 90 % en 58 casos, esperamos que se haya registrado temperatura positiva en 52 de los casos, y efectivamente así habría sido en 52. En un sistema predictivo imperfecto, hemos emitido probabilidad del 90 % en 58 casos, esperamos que se haya registrado temperatura positiva en 52, y realmente se registró en 54. La correspondencia entre probabilidad prevista y frecuencia de observación condicionada definen, así, la llamada *fiabilidad* de un sistema predictivo probabilista. En la Tabla 15.4 se muestra el ejemplo completo. A partir de estas consideraciones puede construirse el llamado *diagrama de fiabilidad* o *de atributos*, con probabilidades previstas en el eje X y las correspondientes frecuencias de observación condicionadas en el eje Y, Figura 15.19. Un sistema perfectamente fiable tendrá su curva de fiabilidad sobre la diagonal, mientras que los sistemas reales tienen su curva usualmente fuera de la diagonal.

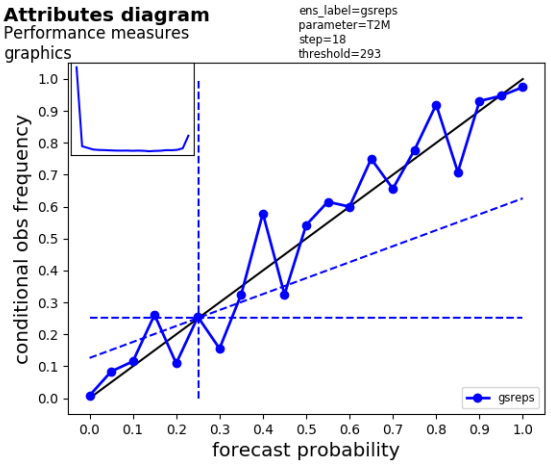

Figura 15.19: Diagrama de atributos correspondiente a los datos de la Tabla 15.4 (ver texto).

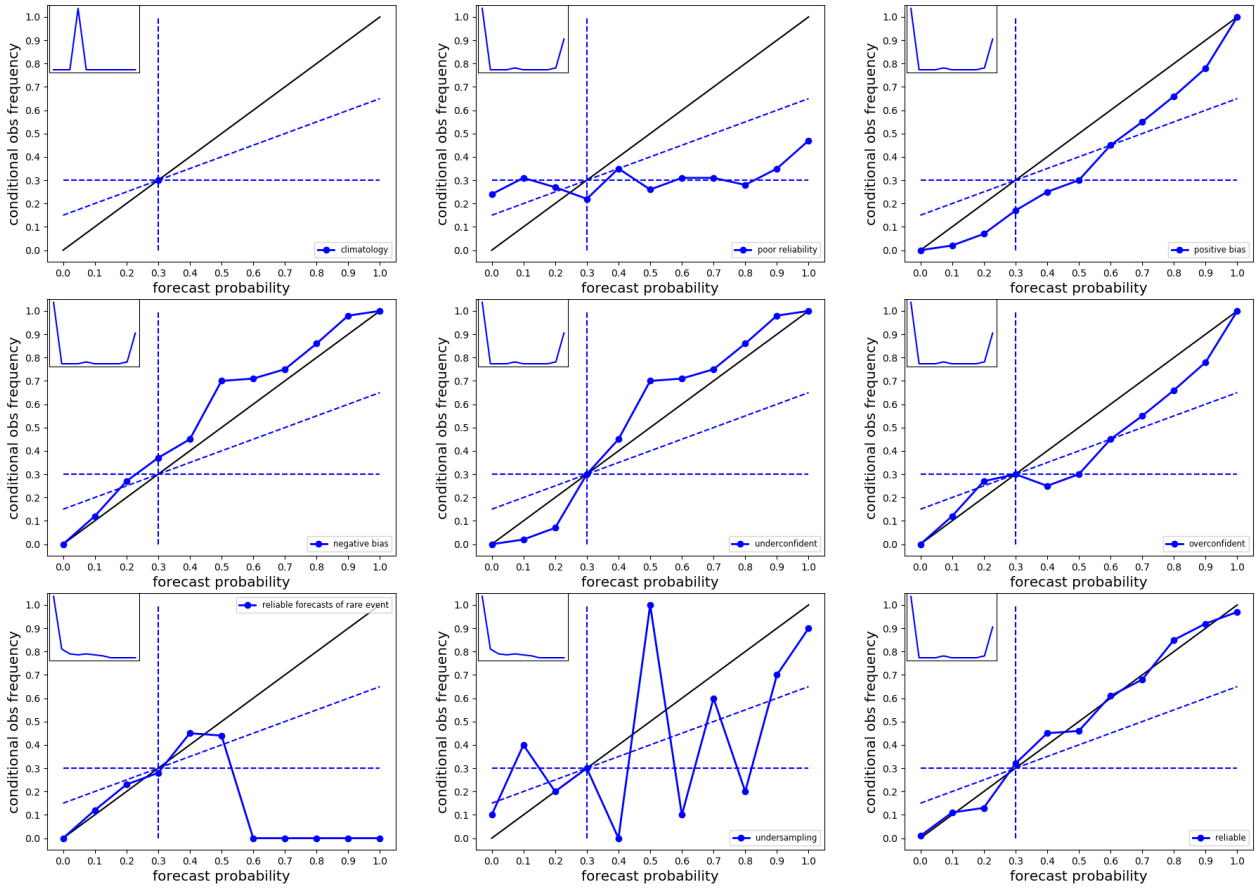

Figura 15.20: Diagramas de atributos correspondientes a respuestas con diferentes grados y matices de fiabilidad. De arriba abajo y de izquierda a derecha, se muestran: climatología muestral, fiabilidad pobre, sesgo positivo, sesgo negativo, sobreconfiado, subconfiado, fiable para un evento raro, submuestreo y fiable (ver texto para descripciones detalladas).

El modo en que la curva de fiabilidad se separa de la diagonal nos informa detalladamente sobre aspectos del sistema predictivo. En la Figura 15.20 se muestra un abanico de posibles matices. La frecuencia de ocurrencia de la climatología muestral usada como predicción (arriba, izquierda) obtiene un diagrama de fiabilidad con un solo punto, tanto la X como la Y en la frecuencia de ocurrencia (así es como se definía); por tanto, es perfectamente fiable y, también, bastante poco útil, en tanto en cuanto tiene muy poca agudeza. Arriba, centro se muestra un ejemplo de fiabilidad pobre, muy plana y alejada de la diagonal (es también un ejemplo de resolución probabilista mínima que veremos más tarde). Arriba en el centro vemos lo que ocurre cuando las predicciones presentan un sesgo positivo: en el espacio probabilista la curva se desplaza hacia abajo; esta forma es muy usual en precipitación. En el centro izquierda algo similar, con sesgo negativo: la curva se desplaza hacia arriba.

En el centro presentamos un ejemplo de predicciones subconfiadas: dan menos probabilidad por encima de la frecuencia de ocurrencia y más por debajo. En el centro derecha el caso contrario, de predicciones sobreconfiadas. Abajo a la izquierda podemos ver un ejemplo de sistema fiable para un evento raro; cuando el evento se da pocas veces, entonces la agudeza se vuelca más en los 0 % y, a partir de los 50 %, 60 % o un intervalo dado, no se emite, ni observa, ningún caso, por lo que la curva colapsa en el cero. Abajo en el centro podemos ver el problema del muestreo escaso (*undersampling*): cuando la muestra es pequeña, ademas de no haber significación estadística, el diagrama es ruidoso, con una curva que da muchos saltos. Por último, abajo a la derecha vemos una verificación con altísima fiabilidad, muy cerca de la diagonal. De un modo aproximado, la pendiente de la curva de fiabilidad es una primera medida cuantitativa de fiabilidad: 1 para fiabilidad perfecta y 0 para utilidad marginal, con utilidades intermedias entre 0 y 1 [26, 45, 48].

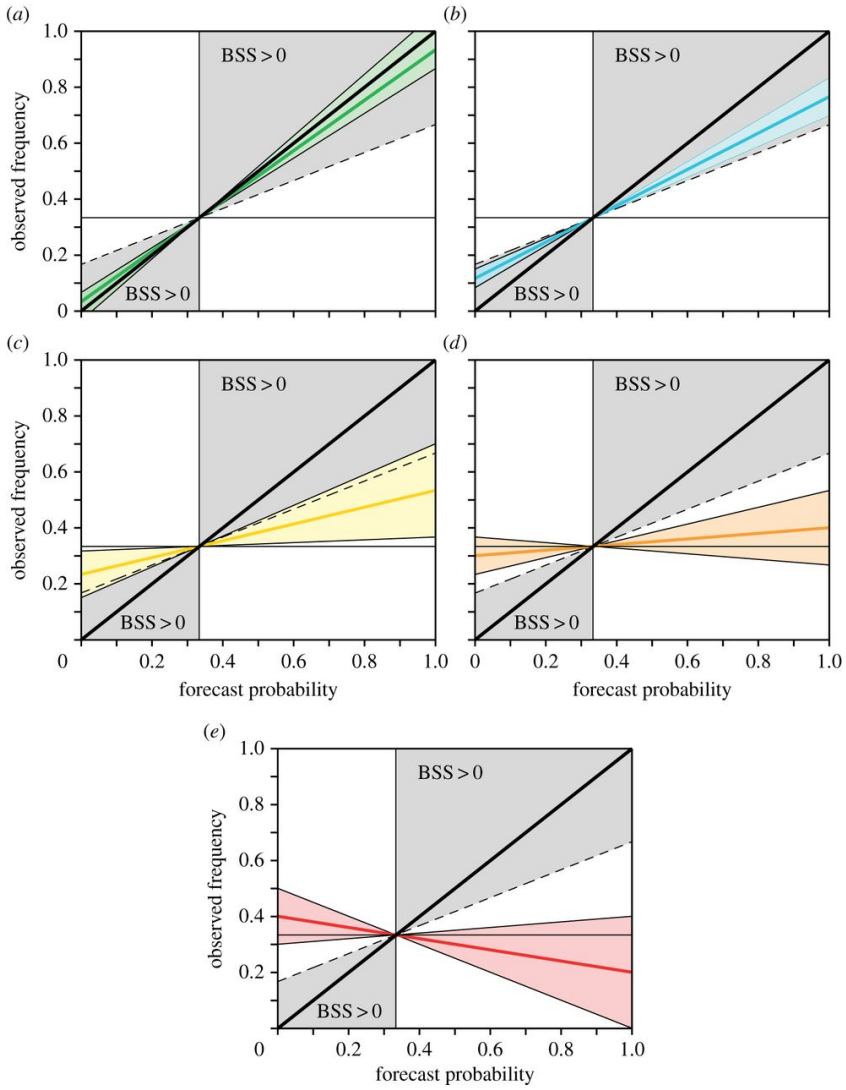

Figura 15.21: Relación entre fiabilidad y habilidad [45] (ver texto). ECMWF.

La relación entre fiabilidad y habilidad (*skill*) puede demostrarse algebraicamente (no se aborda aquí) y tiene una interpretación geométrica sobre el diagrama de fiabilidad. Si trazamos dos rectas, una horizontal y otra vertical, en la abscisa y en la ordenada de la frecuencia de ocurrencia y, después, trazamos la bisectriz de la horizontal mencionada con la diagonal, el diagrama queda dividido en varias regiones, según muestra la Figura 15.21 (a). Con estas características adicionales, el *diagrama de fiabilidad* pasa a llamarse propiamente *diagrama de atributos*. Si la curva queda sobre la región gris, entonces el *BSS* (ver sec. 15.9.8 en la página 232) es positivo y el sistema presenta más skill que la frecuencia de ocurrencia o climatología muestral. Este es el caso en (a), muy fiable y en (b), menos fiable. Si la curva queda en la región blanca, entonces el BSS es negativo y el sistema presenta menos skill que la climatología muestral. Este es el caso en (c), todavía con una fiabilidad útil, en (d), con poca fiabilidad y en (e), con una preocupante fiabilidad por debajo de la horizontal, caso en que se desaconseja usar el sistema predictivo.

Considerando que cuanto más se separa de la diagonal la curva de fiabilidad menos fiable es el sistema, podemos definir una medida cuantitativa de fiabilidad mediante la expresión

$$fiab = \frac{1}{N}\sum_{i=1}^{I} n_i\,(p_i - o_i)^2 \qquad (15.9)$$

donde N es el número total de casos, I es el número de intervalos de probabilidad (a menudo 10), n_i es el número de casos en el intervalo i, p_i es la probabilidad en el intervalo i (e. g. 90 % ó 0.9), o_i es la frecuencia observada del evento en el intervalo i (e. g. 80 %).

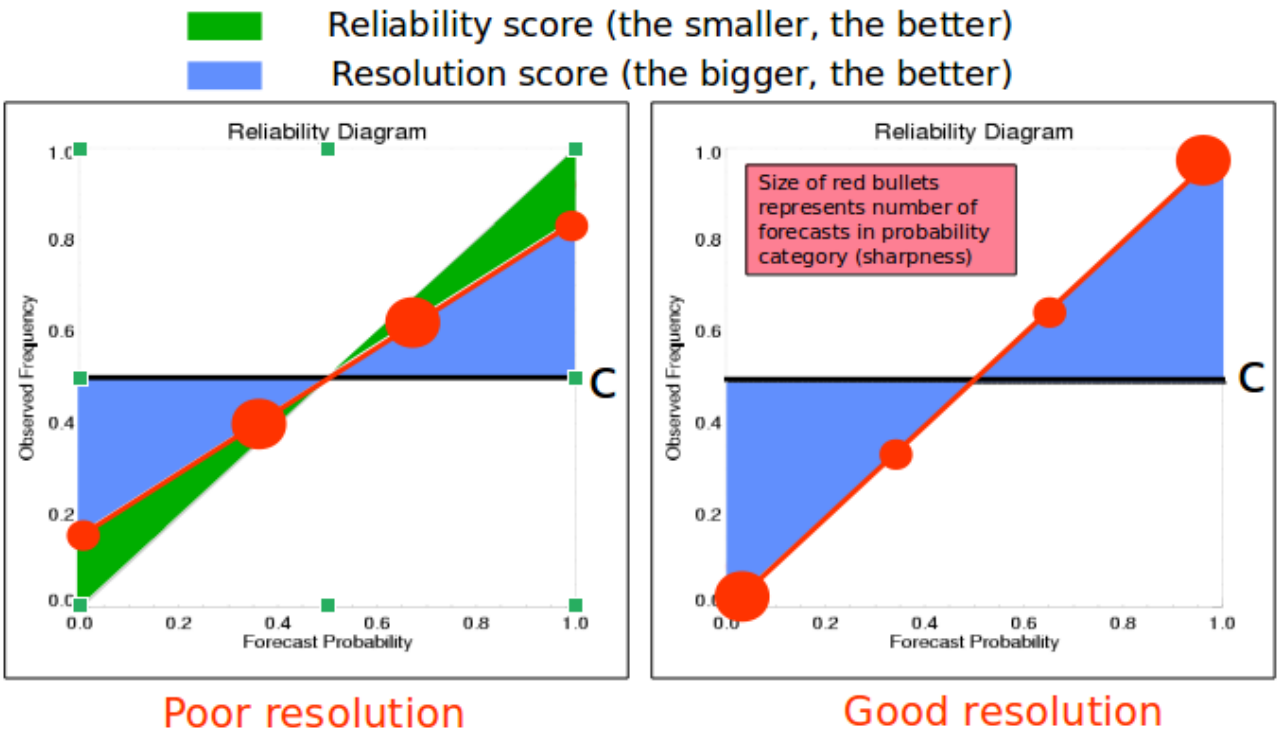

Figura 15.22: Ilustración teórica de las propiedades fiabilidad y resolución probabilista. En estos diagramas el histograma de agudeza se sustituye por el artificio de dar distintos tamaños a los círculos que representan cada puinto en el diagrama. A la izquierda se muestra una curva sobreconfiada, con una agudeza abultada en los intervalos intermedios (baja), con resolución (distancia a la recta horizontal «sin habilidad») relativamente baja. A la derecha vemos una curva con fiabilidad perfecta, mayor resolución y agudeza volcada en los intervalos extremos (alta). ECMWF.

15.9.5 Resolución

En el diagrama de atributos, la recta horizontal cuya ordenada es la frecuencia de ocurrencia representa la predicción teórica más sencilla, el uso de la climatología muestral ya contado anteriormente. Esa predicción es el «listón más bajo» en términos de habilidad, i. e. una predicción con pretensiones de aportar habilidad tiene que mejorar la «habilidad nula» de la climatología muestral, que sólo emite predicciones probabilistas fijas, cuyo valor es precisamente, la frecuencia de ocurrencia. Cualquier predicción que mejore esta pauta tiene que «resolver» en diferentes intervalos de probabilidad, teniendo en cuenta las observaciones.

La separación de la curva de fiabilidad con respecto a esta recta horizontal define, en este espacio de probabilidades, la propiedad que denominamos resolución. La *resolución probabilista* es la capacidad de un sistema predictivo de emitir probabilidades en distintos intervalos, con una «resolución» mejor que la de la climatología muestral.

Considerando que cuanto más se separa la curva de fiabilidad de la recta horizontal más resolución presenta el sistema, podemos definir una medida cuantitativa de resolución mediante la expresión

$$reso = \frac{1}{N}\sum_{i=1}^{I} n_i\,(s-o_i)^2 \qquad (15.10)$$

donde N es el número total de casos, I es el número de intervalos de probabilidad (a menudo 10), n_i es el número de casos en el intervalo i, s es la frecuencia climatológica de ocurrencia observada en toda la muestra (e. g. 30 % ó 0.3), o_i es la frecuencia observada del evento en el intervalo i (e. g. 80 %).

En la Figura 15.23 en la página siguiente se muestra un diagrama de atributos con varias curvas de fiabilidad de casos reales.

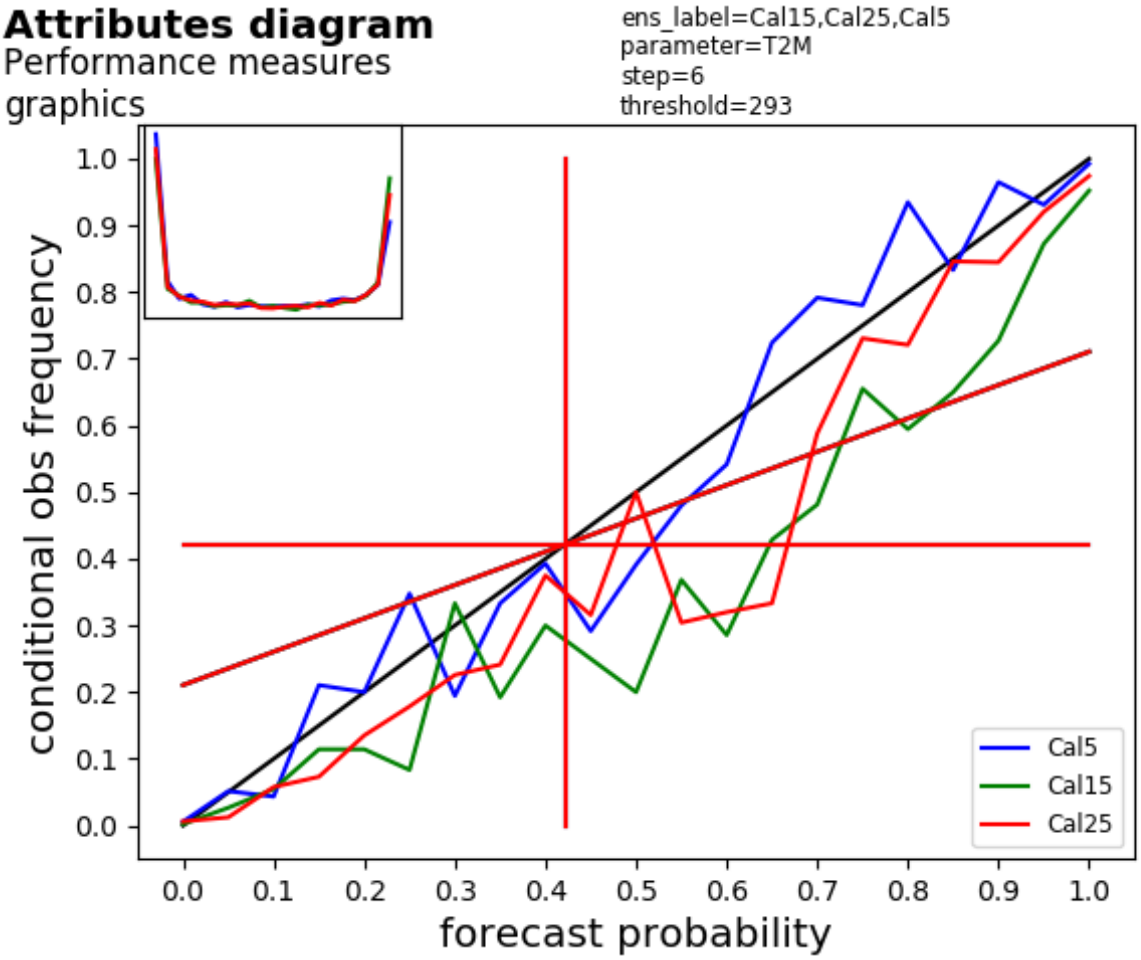

Figura 15.23: Diagrama de atributos con curvas de fiabilidad (parte central) e histogramas de agudeza (esquina superior izquierda) de distintas predicciones, correspondientes a distintas calibraciones del sistema AEMET-γSREPS, considerando el evento de temperatura por encima del umbral fijo de 293 K. Puede apreciarse apreciándose que Cal5 está más cerca de la diagonal y por ende es, en general, más fiable y con mejor resolución. Cal15 está claramente bajo la diagonal, con pendiente inferior, menos fiable y con sesgo positivo. Cal25 está en general bajo la diagonal pero presenta mejor pendiente que Cal15. Los histogramas de agudeza muestran comportamientos similares, aunque en el intervalo de mayor probabilidad (90-100%) Cal15 muestra más agudeza, Cal25 por detrás y Cal5 el último.

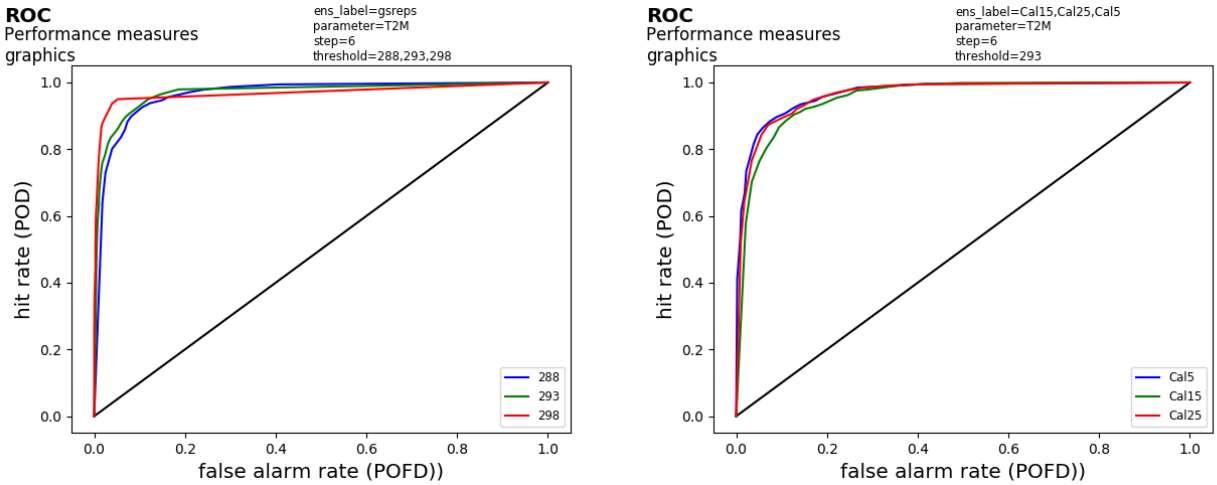

Figura 15.24: Curvas ROC de distintas predicciones. Izquierda: correspondientes al sistema AEMET-γSREPS, probabilidades de temperatura por encima de los umbrales de 288 (azul), 293 (verde) y 298 K (rojo); para 293 (20 ºC) hay un rendimiento ligeramente mejor que para 288 (15 ºC), mientras que para 298 K (25ºC), al tratarse de un umbral superior el evento es más raro, volcando los puntos de la curva sobre el lado izquierdo y bajando algo el área bajo la curva. Derecha: distintas calibraciones del mismo sistema AEMET-γSREPS, para el umbral fijo de 293 K, apreciándose que Cal5 da un resultado muy ligeramente mejor que Cal15, a su vez algo mejor que Cal25.

15.9.6 Discriminación y curva ROC

Una medida complementaria de rendimiento de un SPC es la *discriminación* o capacidad de un SPC para distinguir entre la ocurrencia o no ocurrencia de un evento binario (normalmente superación de un umbral), dadas las observaciones. Este concepto procede de la Teoría de Detección de la Señal y está relacionado con la tasa de aciertos, hit rate (H) y la tasa de falsas alarmas, false alarm rate (F), para una frecuencia de ocurrencia dada [7, 26, 27, 28].

Una medida usual de discriminación es el área bajo la llamada *curva ROC* [24, 29], curva en la que se dibuja F frente a H (Figura 15.24). La climatología muestral usada como predicción obtendría un área de 0.5 (sin skill), mientras que una predicción perfecta obtendría un área de 1.0. A menudo se usa la medida llamada ROC Skill Area (RSA), dada por $RSA = 2A - 1$ que brinda valores entre -1 y +1: +1 para la predicción perfecta, 0 para predicciones sin habilidad y -1 para una predicción potencialmente perfecta después de un adecuado proceso de calibración.

La discriminación está relacionada con la resolución probabilista, pero no miden exactamente la misma propiedad y, especialmente cuando hay incertidumbre observacional presente pueden mostrar comportamientos diferentes indicativos. La medida BSS es potencialmente insensible a fenómenos extremos y RSA no lo es [20], aunque RSA puede ser insensible a algunos tipos de sesgos predictivos[29].

15.9.7 Resumen de propiedades

Frecuencia de ocurrencia de observación: frecuencia con la que el evento ocurrió en el periodo.

Agudeza: frecuencia de cada probabilidad prevista emitida, sin atender a la observación.

Fiabilidad: frecuencias de observación condicionadas con respecto a probabilidades previstas.

Resolución probabilista: capacidad para emitir probabilidades fiables que se distingan de la climatología muestral.

Discriminación: capacidad para distinguir la ocurrencia o no ocurrencia del evento, dadas las observaciones.

15.9.8 Conexión Brier - propiedades

El índice de BRIER (BRIER *score*) puede descomponerse [26, 28, 48] según la ecuación 15.12, en la que podemos reconocer las medidas cuantitativas de fiabilidad y resolución que habíamos definido en el contexto de las distribución conjunta. La tercera componente resulta ser el índice de BRIER que correspondería a la climatología muestral como sistema predictivo, por lo que se denomina a menudo BS_{clim}, aunque es más acertado llamarla componente de incertidumbre BS_{ince}, pues tiene que ver con la incertidumbre asociada a las predicciones en el periodo, de hecho es la varianza de la frecuencia de ocurrencia. Esta descomposición del índice de BRIER conecta, así, los dos enfoques para eventos binarios: el de BRIER y el de distribución conjunta y llamaremos a las distintas componentes según su significado:

$$BS = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n)^2 = \\ = \underbrace{\frac{1}{N}\sum_{i=1}^{I} n_i (p_i - o_i)^2}_{BS_{fiab}} - \underbrace{\frac{1}{N}\sum_{i=1}^{I} n_i (s - o_i)^2}_{BS_{reso}} + \underbrace{s(1-s)}_{BS_{ince}} \tag{15.11}$$

En resumen:

$$BS = BS_{fiab} - BS_{reso} + BS_{ince} \tag{15.12}$$

Dado que el índice de BRIER es una magnitud de error y cuanto más cercano a cero mejor es el sistema, la componente de fiabilidad será mejor cuanto más pequeña, la componente de resolución lo contrario y la de incertidumbre igual que la de fiabilidad.

Se define, también un BRIER *skill score* (*BSS*), usualmente tomando como skill de referencia el de la climatología muestral, es decir, lo que hemos denominado BS_{ince}, de modo que:

$$BSS = \frac{BS - BS_{ince}}{BS_{perfecto} - BS_{ince}} = \frac{BS - BS_{ince}}{0 - BS_{ince}} = 1 - \frac{BS}{BS_{ince}} \tag{15.13}$$

Sustituyendo BS por su descomposición y ordenando todos los términos, resulta la descomposición del BSS:

$$BSS = 1 - BSS_{fiab} - BSS_{reso} \tag{15.14}$$

15.9.9 Ranked y Continuous ranked probability score (CRPS)

El *BS* y el *BSS* van ligados a la superación de un umbral fijo de la variable meteorológica. Si nos interesa cubrir un número de umbrales o, incluso, todo el abanico de posibles umbrales, entonces definimos el Ranked probability score (*RPS* [26, 28]) o el Continuous ranked probability score (*CRPS* [7, 9, 25, 48]) mediante una suma, o bien una integral extendida a todos los valores posibles del umbral:

$$RPS = \frac{1}{M-1}\sum_{m=1}^{M}\left[\left(\sum_{k=1}^{m} p_k\right) - \left(\sum_{k=1}^{m} o_k\right)\right]^2$$

$$CRPS = \int_{-\infty}^{\infty}\left(P_f(x) - P_0(x)\right)^2 dx \tag{15.15}$$

Existen, también, los correspondientes *skill scores* *RPSS* y *CRPSS*. Mientras que los scores tipo error están orientados negativamente (cuanto más cercanos a cero, mejor), los skill scores están construidos para estar orientados positivamente: mejor cuanto más grande. Así, en las gráficas de evolución en el tiempo, con los años puede observarse si las curvas «crecen», sin necesitar conocer el significado profundo de la métrica.

Por un lado, el BS o el BSS se utilizan para evaluaciones en umbrales específicos, típicamente en contextos de predicción operativa donde los umbrales de importancia son fijos, e. g. precipitación por encima de 1, 5, 10 y 20 mm. Por otro lado, el *RPS* o el *CRPS* se utilizan más como métricas resumen (*summary measures*), cuando interesa sintetizar el rendimiento del sistema en pocas cantidades. La Figura 15.25 en la página siguiente ilustra el uso del CRPSS para monitorizar el progreso del ECENS a lo largo de dos décadas, en términos de límite de predecibilidad.

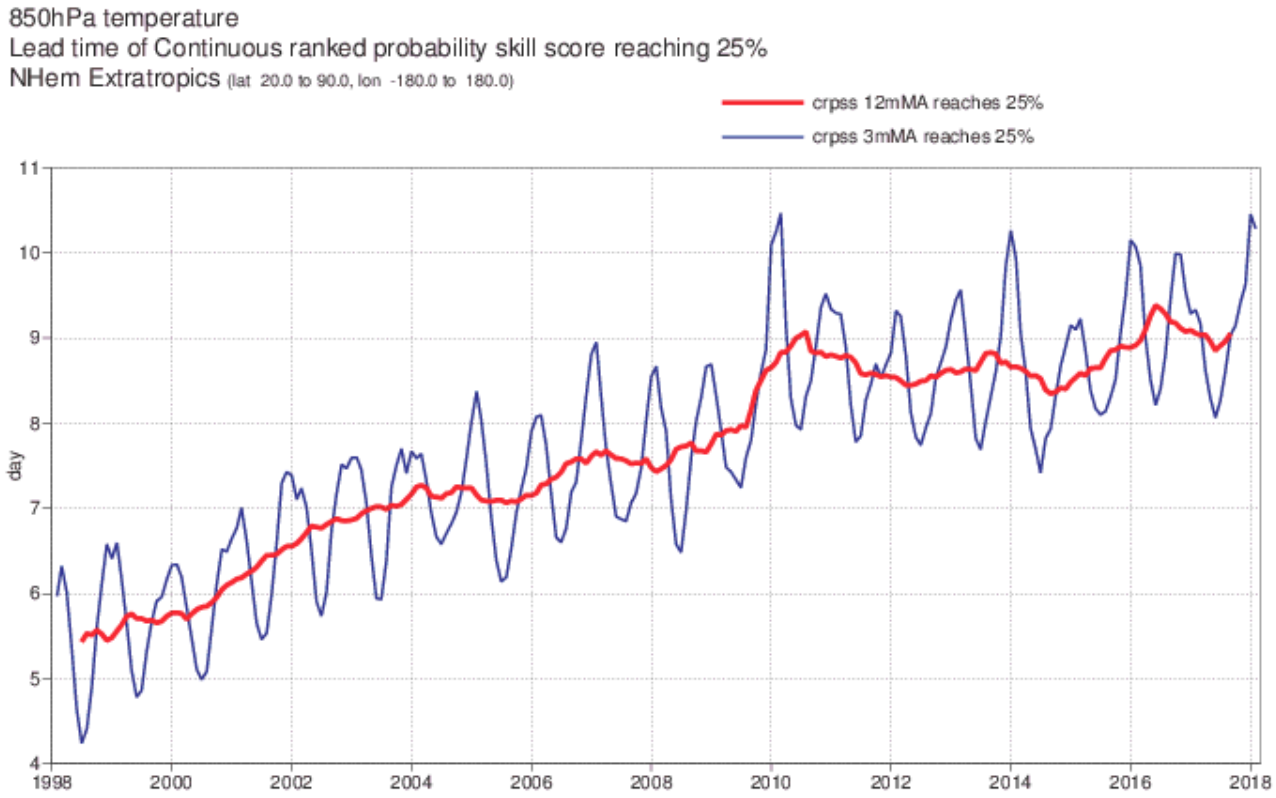

Figura 15.25: Serie temporal de calidad de predicciones de temperatura en 850 hPa del ECENS, en la franja extratropical norte. La calidad se mide por límite de predecibilidad: el alcance predictivo en que el CRPSS desciende por debajo del 25 % (hasta entonces había estado por encima y la calidad de la predicción se había mantenido «buena»). La curva azul es una media móvil mensual (para eliminar señales de alta frecuencia, i. e. variabilidades diarias), mientras que la roja es una media móvil trimestral por consiguiente aún más suave. Puede observarse que en el año 2 000 las predicciones de ECENS eran buenas hasta el día 5.8 aproximadamente, alcanzando o superando en la actualidad el D+9: se han ganado algo más de tres días de predecibilidad en veinte años, día y medio por década. ECMWF.

Tabla contingencia		
	SI ocurre	NO ocurre
SI previsto	a	b
NO previsto	c	d

Matriz costes		
	SI ocurre	NO ocurre
SI actuar	C	C
NO actuar	L	0

Figura 15.26: Los dos aspectos principales del análisis económico de una predicción desglosados en detalle. Por un lado, la información meteorológica y del sistema predictivo está contenida en la denominada tabla de contingencia (izquierda), tabla de doble entrada con contabilidad sobre lso casos disversos en que el fenómeno meteorológico (e. g. lluvia) se predijo o no y tuvo lugar o no. Por otro lado, la información económica del usuario se detalla en la denominada matriz de costes (derecha), también de doble entrada, tgeniendo en cuenta si el usuario toma o no acción preventiva, acarreando un coste diario C o provocando una pérdida diaria L en auqellos días en que no se tomó acción preventiva y el fenómeno tuvo lugar (ver texto).

15.10 Valor (económico) relativo

Las predicciones de cualquier naturaleza deben ayudar al usuario en la toma de decisiones. Como hemos mostrado previamente, la *teoría de la detección de la señal* nos permite mostrar la respuesta de un sistema predictivo frente a eventos binarios (ocurre / no ocurre), con dos parámetros predictivos del sistema: la tasa de aciertos (hit rate, H) y la de falsas alarmas (false alarm rate, F) y con un parámetro que depende sólo de la ocurrencia del fenómeno: la frecuencia de ocurrencia observada (base rate s); hablamos en ese contexto de una propiedad llamada discriminación y de unos umbrales de probabilidad óptimos. Hasta aquí la información manejada es de carácter estrictamente meteorológico.

Para permitir entrar en juego a parámetros que dependen de las características del usuario de las predicciones, podemos aplicar ahora elementos de *teoría de la decisión*. Introduciremos los costes C y las pérdidas L relacionados con las implicaciones socioeconómicas del fenómeno, sabiendo que la ocurrencia de dicho fenómeno tiene una frecuencia s.

Los costes C son los costes económicos, o de otro tipo cuantificable, que acarrea una acción preventiva relacionada con el fenómeno.

Las pérdidas L representan las pérdidas económicas, o de otro tipo cuantificable, que tendrían lugar si el fenómeno ocurre y no hemos tomado las acciones preventivas correspondientes.

Ilustremos estos conceptos con un ejemplo sencillo: una bar con terraza al aire libre.

Ejemplo: terraza al aire libre. En la construcción más sencilla posible, imaginemos un bar que puede disponer de terraza al aire libre en la temporada de verano en el Cantábrico, área donde no siempre está garantizado el tiempo estable. El evento o fenómeno meteorológico de interés es la ocurrencia de lluvia (llueve o no llueve). El periodo de interés puede ser de algo más de tres meses, digamos 100 días. La frecuencia de ocurrencia puede tener, por ejemplo, un valor de 0.3, el 30 % de los días tiene lugar la lluvia, es decir llueve 30 días y no llueve 70 días. Los dueños del bar calculan los costes de tomar acción preventiva (recoger mesas, desplegar toldos, etc.) en unos 20 euros al día y, por otro lado, las pérdidas ocasionadas por la lluvia (ausencia de clientela) en unos 200 euros cada vez que llueva sin tener toldos. En este caso sencillo la probabilidad de lluvia, contando con la climatología muestral, es de 0.3. Calculamos el coste de tomar acción preventiva $C_{tot} = 100 \times 20 = 2000$ euros y las pérdidas esperadas en caso de no tomarla $L_{tot} = (a+c)L = 30 \times 200 = 6000$ euros. Y decidimos que merece la pena gastar 2 000 euros para no perder 6 000. La comparación, en términos matemáticos, puede expresarse como:

$$\begin{aligned} &C_{tot} < L_{tot} \Rightarrow NC < (a+c)L \Rightarrow \\ &\Rightarrow C < \tfrac{a+c}{N} L \Rightarrow C < sL \Rightarrow \tfrac{C}{L} < s \end{aligned} \tag{15.16}$$

Es decir, para que los costes totales sean inferiores a las pérdidas, la proporción $\frac{C}{L}$ debe ser menor a la frecuencia de ocurrencia del evento. En el caso de la terraza $\frac{20}{200} = 0{,}1 < 0{,}3$, por tanto es rentable tomar la decisión conociendo la climatología muestral y la relación coste-pérdida del usuario. Pero no hemos utilizado información de la calidad del sistema predictivo. para ello hay que refinar la idea entrando en detalle.

Refinamiento y detalle de la idea. La información meteorológica puede sintetizarse en una tabla de contingencia (Figura 15.26 en la página anterior izquierda), tabla de doble entrada. Dos filas desglosan si el evento está previsto o no lo está. Dos columnas desglosan, a su vez, si el evento tuvo lugar o no. De modo que podemos contabilizar el número de veces que el evento:

- Estaba previsto y ocurrió: aciertos a.
- Estaba previsto pero no ocurrió: falsas alarmas b.
- No estaba previsto pero ocurrió: errores c.
- Ni estaba previsto ni ocurrió: negativos correctos d.

Con esta información desglosada, la tasa de aciertos es la frecuencia con la que está previsto el evento dentro del total de casos en que tuvo lugar, es decir $H = \frac{a}{a+c}$. Por otro lado, la tasa de falsas alarmas es la frecuencia con la que se predijo, dentro del total de casos en que no ocurrió, es decir $F = \frac{b}{b+d}$. La frecuencia de observación u ocurrencia del fenómeno es $s = \frac{a+c}{a+b+c+d} = \frac{a+c}{N}$. Al conjunto de información relativo al periodo de interés lo llamamos *climatología muestral*.

El usuario de la predicción se caracteriza, en este esquema, mediante la llamada matriz de costes (*expenses matrix* en inglés, Figura 15.26 en la página anterior derecha), también de doble entrada. Dos filas desglosan si se han tomado medidas preventivas con su correspondiente coste C o, si por el contrario, no se han tomado, con la consiguiente pérdidas L. Dos columnas desglosan si el evento ha tenido o no lugar. De modo que podemos detallar los costes y pérdidas del siguiente modo:

- Se actuó preventivamente y ocurrió: ha costado C.
- Se actuó preventivamente pero no ocurrió: ha costado C.
- No se actuó preventivamente y ocurrió: ha provocado pérdidas L.
- No se actuó preventivamente pero no ocurrió: no hay pérdidas.

El estudio económico detallado se realiza entonces cruzando las dos tablas: la tabla de contingencia, con la información meteorológica y la matriz de costes con la información económica del usuario. Se pueden dar varios casos hipotéticos:

Si imaginamos una *predicción perfecta*, entonces el coste total será el coste diario multiplicado por el número de días en que tiene lugar el evento: $E_p = \frac{a+c}{N}C = sC$. Si asumimos, de un modo más realista, que la predicción no es perfecta, entonces el coste total será el coste diario multiplicado por el número de aciertos, más el coste diario multiplicado por el número de falsas alarmas, más la pérdida diaria multiplicada por el número de errores: $E_f = aC + bC + cL$. Finalmente, si pudiésemos disponer de una bola de cristal con la tabla de contingencia, es decir, información de la *climatología muestral*, antes del periodo, podríamos usar la climatología muestral para dar una predicción probabilista sencilla: todos los días daríamos una probabilidad s de ocurrencia del fenómeno. Estadísticamente hablando, es una predicción perfectamente fiable, porque una vez transcurrido el periodo el fenómeno ha tenido lugar con frecuencia s y lo hemos previsto asimismo con frecuencia s. En ese caso, económicamente, habremos de asumir bien un coste C o bien un coste $\frac{a+c}{N}L = sL$, tomando el mínimo entre ambas cantidades: $E_c = min(C, sL)$. Definimos el valor económico de la predicción en el periodo de interés como la reducción económica que nos aporta el uso de la predicción en comparación con el hipotético uso de una predicción perfecta, en ambos casos con respecto a la climatología muestral (bola de cristal) como predicción.

Valor económico relativo. La reducción económica usando nuestro sistema con respecto a la climatología muestral es $E_f - E_c$,y la reducción económica de una predicción perfecta sería $E_p - E_c$. De modo que el valor económico relativo, es decir, la relación entre nuestra reducción económica y la de la predicción perfecta, vendrá dado por:

$$RV = \frac{E_f - E_c}{E_p - E_c} \tag{15.17}$$

Introduciendo las expresiones previas para cada una de las cantidades, resulta la expresión:

$$RV = \begin{cases} (1-F) - \left(\frac{1-C/L}{C/L}\right)\left(\frac{s}{1-s}\right)(1-H) & \text{si } \frac{C}{L} > s \\ H - \left(\frac{C/L}{1-C/L}\right)\left(\frac{1-s}{s}\right)F & \text{si } \frac{C}{L} < s \end{cases} \tag{15.18}$$

Como vemos, el valor económico relativo depende, por un lado, de la calidad de la predicción, en términos de su tasa de aciertos H y de su tasa de falsas alarmas F, además de la propia frecuencia de ocurrencia del fenómeno s y, por otro lado, de las características económicas del usuario, en terminos de sus costes C a la hora de tomar acción preventiva frente al fenómeno y de las pérdidas L ocasionadas en caso de no tomarla.

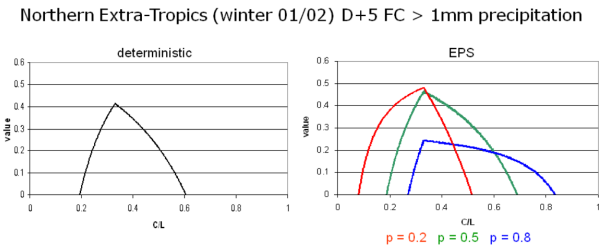

Figura 15.27: Curvas de valor (económico) relativo correspondientes al modelo ECHRES (izquierda) y al SPC ECENS (derecha), como ejemplo de curvas de valor. Aunque ECENS consta de 51 miembros y por tanto podrían dibujarse hasta 51 curvas, aquí se pintan solamente 3 curvas, a título ilustrativo, correspondientes a probabilidades de p=*0.2 (naranja),* p=*0.5 (verde) y* p=*0.8 (azul). El eje horizontal representa la proporción coste/pérdida C/L y el eje vertical el valor económico relativo RV. Obsérvese que cada intervalo de probabilidad disfruta de un RV máximo diferente, pero todos ellos correspondientes a un mismo C/L, en particular $C/L = s$. ECMWF.*

Un sistema predictivo determinista, e. g. el modelo ECHRES tendrá asociada una curva de valor relativo que viene dada por la expresión 15.18 en la página anterior. Su representación es una gráfica con C/L en el eje horizontal, *RV* en el eje vertical y una curva en forma de montículo, más o menos simétrica y con la cúspide más o menos centrada (Figura 15.27 izquierda).

En un SPC o, en general, un sistema de predicción probabilista cualquiera, la calidad y el rendimiento suelen medirse con una partición en el espacio de las probabilidades. Esa partición es óptima en tantos intervalos como miembros tiene el sistema, aunque a menudo, por simplicidad, suele realizarse en 10 intervalos de amplitud 0.1. Se obtienen así medidas de calidad en cada intervalo, que sirven para construir medidas más complejas. Así se hacía con la curva ROC, conjunto de pares (H,F) y así lo hacemos ahora con el valor relativo *RV*: un conjunto de curvas $RV(C/L)$, una para cada intervalo de probabilidad. Como ejemplo, en la Figura 15.27 derecha, se pintan tres curvas de *RV* del ECENS correspondientes a probabilidades de 0.2, 0.5 y 0.8. El valor máximo se da en $C/L = s$ y, en teoría, cuando un sistema fiable (sec. 15.9.4 en la página 226) predice un evento con probabilidad p, entonces todos los usuarios con $C/L < p$ deberían actuar.

Un ejemplo más completo se muestra en la Figura 15.28, con curvas de valor (económico) relativo correspondientes al sistema AEMET-SREPS, con 16 curvas, una curva de valor para cada intervalo de probabilidad. El valor relativo neto del SPC, interpretado gráficamente, es la envolvente superior a todas las curvas, es decir, el valor máximo para cada valor de C/L.

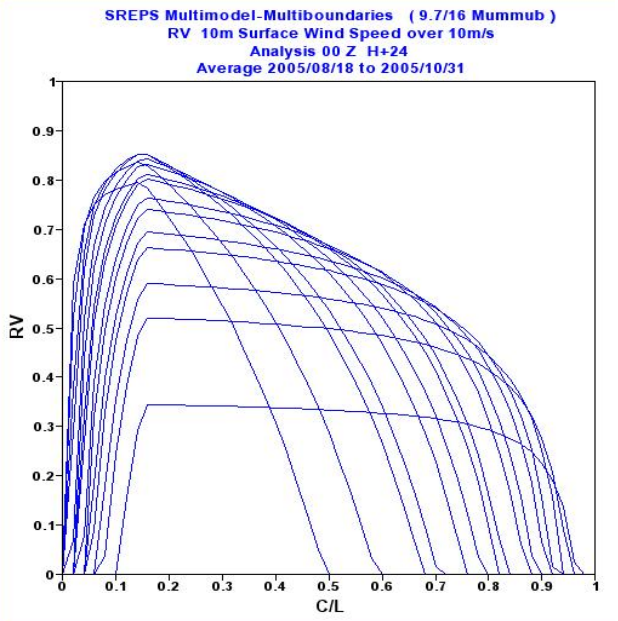

Figura 15.28: Curvas de valor (económico) relativo correspondientes al sistema AEMET-SREPS, predicciones probabilistas H+24 de viento a 10 m superior a 10 m/s en un periodo de varios meses de 2005. En ese periodo el sistema constaba de 16 miembros, de modo que la partición natural del espacio de probabilidades es en 16 intervalos, con una curva de valor para cada uno de ellos. El valor relativo es la envolvente superior a todas las curvas, es decir, el valor máximo para cada valor de C/L.

Esta construcción cuantitativa del valor económico nos permite, entre otros muchos propósitos, hacer una comparación directa entre un modelo determinista y un SPC. Antes de mostrar la idea, Es importante subrayar que, desde un punto de vista pragmático e histórico, cada tipo de sistema ha tenido un nicho ecológico distinto, sin tener necesariamente por qué competir. Aunque en muchas ocasiones puede haberse caído en una competitividad innecesaria (como muchas veces ocurre). Al contrario, los modelos deterministas son elementos esenciales para los SPC, de modo que no hay competencia sino integración. La comparación puede hacerse, eso sí, con propósitos educativos, ilustrativos e, incluso, económicos, literalmente para comparar el valor económico relativo de usar un SPC o de usar un modelo determinista en un contexto como podría ser la energía eólica o la solar. Puede así, por ejemplo, convencerse a usuarios veteranos reacios a

adentrarse en el mundo de las probabilidades. La Figura 15.29 muestra una comparación de este tipo, valor de un SPC frente a un modelo determinista, en este caso entre el ECENS en conjunto y el ECHRES, modelo determinista de mayor resolución espacial. El uso del ECENS como conjunto es mucho más valioso que el uso de cada uno de sus miembros individualmente.

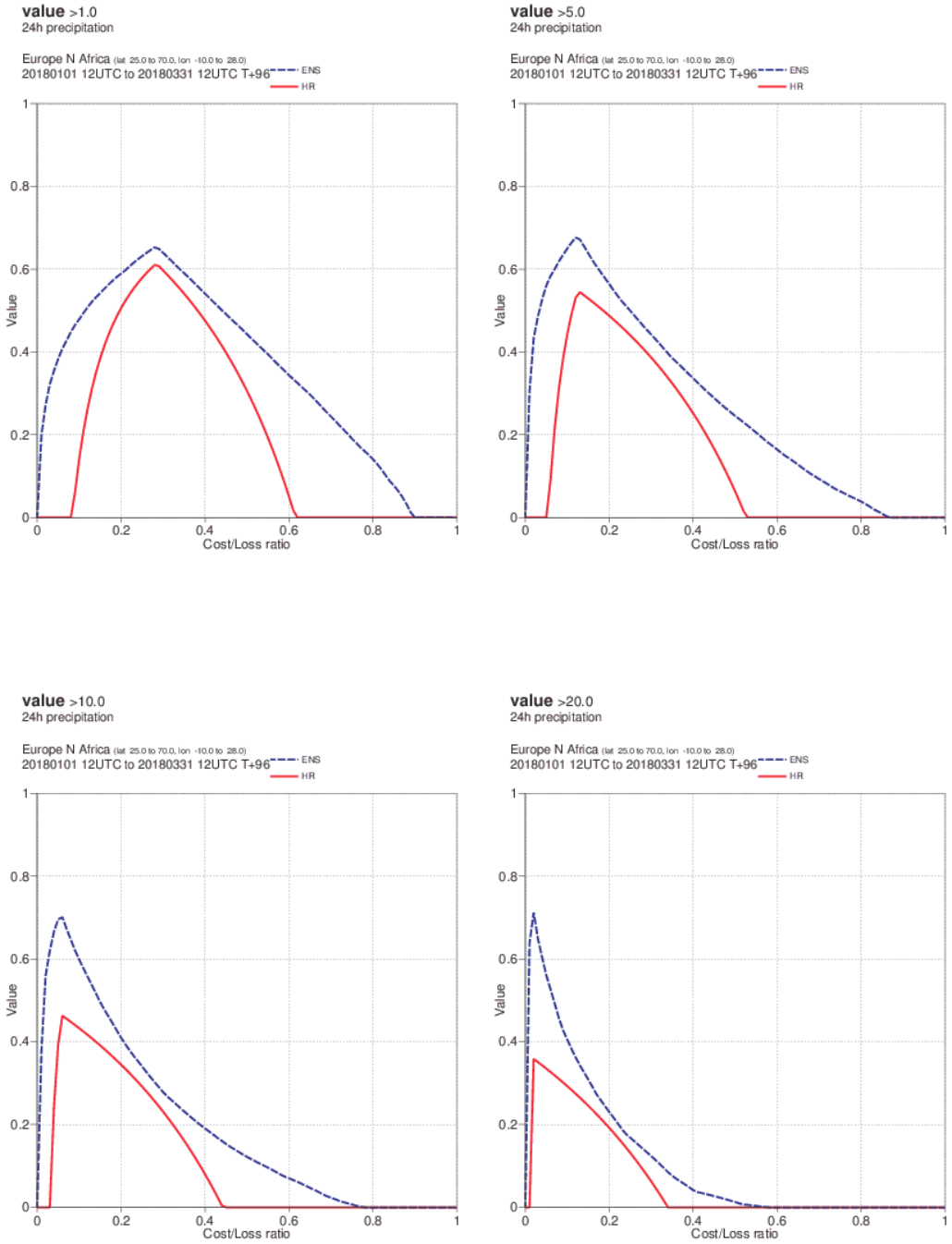

Figura 15.29: Envolventes de valor (económico) relativo correspondientes al ECENS (azul) y curvas de valor relativo del ECHRES, como modelo determinista (rojo), para predicciones de precipitación superior a umbrales de 1, 5, 10 y 20 mm diarios para un alcance de D+4 en la franja extratropical norte del planeta durante los 3 primeros meses de 2018. A medida que el umbral crece, se hace más pequeña la frecuencia de ocurrencia (posición horizontal del máximo). Las curvas azules están por encima de las rojas: el uso del ECENS como conjunto es mucho más valioso que el uso de uno de sus miembros individualmente. ECMWF.

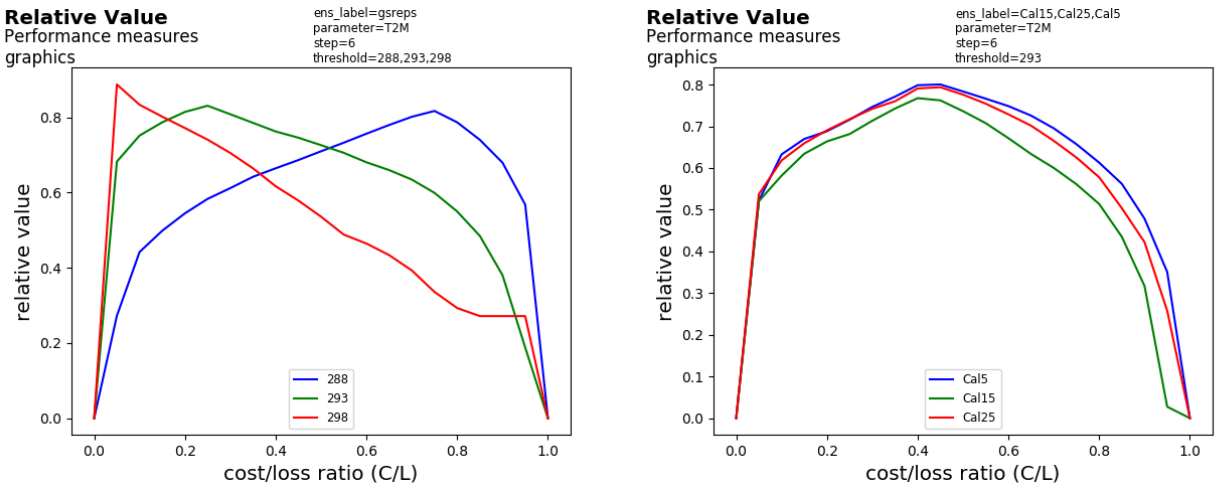

Figura 15.30: Envolventes de valor (económico) relativo correspondientes al sistema AEMET-γSREPS. Izquierda: eventos de temperatura por encima de los umbrales de 288 (azul), 293 (verde) y 298 K (rojo); obsérvese que el valor máximo para cada umbral se da en proporciones C/L más pequeñas según el umbral va creciendo. Derecha: distintas calibraciones del sistema para umbral fijo de 293 K, con valores relativos Cal5 superior a Cal15, a su vez superior a Cal25; el valor máximo se encuentra en una proporción C/L intermedia.

El uso de envolventes de valor económico relativo es sistemático para comparar valores de distintos SPC, o de distintas calibraciones de un mismo SPC o, también, distintos periodos de predicción, alcances, fechas o umbrales de superación en la variable. La Figura 15.30 muestra envolventes de valor económico relativo correspondientes a predicciones de temperatura del sistema AEMET-γSREPS (cap. 22 en la página 333), tanto para distintos umbrales (izquierda), como para diferentes procesos de calibración de esas predicciones (derecha). Las envolventes de la izquierda son resultados del mismo sistema para eventos distintos umbrales, cada uno con una climatología muestral diferente (frecuencia de ocurrencia distinta), lo que se revela en los máximos valores en posiciones distintas. La de la derecha, sin embargo, son resultados de distintas calibraciones, al fin y al cabo, como distintos sistemas, respondiendo a un mismo evento, una misma climatología muestral; de ahí que los valores máximos se producen todos en la misma posición.

15.11 Conclusiones

Ni los modelos atmosféricos ni los SPC son perfectos. Resulta necesario evaluar su calidad, comparándolo con las observaciones de la atmośfera y, asimismo, su valor, contrastando su utilidad para el usuario, especialmente a la hora de tomar decisiones. Estos aspectos conforman la parte de la meteorología y la climatología que podemos denominar verificación. Gracias a la verificación, podemos resaltar puntos fuertes y débiles de los modelos, evaluar su calidad y su valor, mejorar su uso en la predicción operativa, ayudando tanto a predictores como a investigadores y desarrolladores.

La evaluación de un modelo numérico lleva consigo un alto nivel de complejidad, con numerosos problemas técnicos, de interpretación, etc. Es necesario un marco estadístico para dar significación a una serie de propiedades y métricas (también llamadas índices o, usando terminología en inglés, *scores*) que caracterizan la comparación de grandes cantidades de predicciones y observaciones: error medio, error cuadrático medio, correlación de anomalías, etc.

La pregunta de la sociedad ¿hasta cuántos días son fiables las predicciones del tiempo? es, en principio, difícil de responder. Sabiendo que la predecibilidad es dependiente del flujo, esa sería la mejor respuesta: depende de la situación. Pero la sociedad exige cifras para poder hacer planificaciones. La verificación ofrece aquí una respuesta razonable, ligada a lo que denominamos límite de predecibilidad, que es el alcance predictivo máximo en el que la calidad de la predicción, según una métrica establecida, se mantiene por encima de un umbral prefijado. Así, este límite tiene un significado preciso para los científicos y técnicos y, a su vez, lo tiene para la sociedad: con respecto a la década de los 80, hemos ganado algo más de 3 días de límite de predecibilidad, las predicciones son fiables hasta el día más cinco y medio.

Los métodos clásicos de verificación han topado con

limitaciones importantes, algunas de las cuales abren nuevas líneas de investigación:

- Entre la agrupación de suficientes datos para conseguir significación estadística y la estratificación de los mismos para verificar muestras climatológicamente homogéneas, es difícil encontrar un punto intermedio de compromiso.
- El error observacional no se ha tenido tradicionalmente en cuenta y puede influir considerablemente en la verificación.
- Tampoco se ha tenido tradicionalmente en cuenta el error en el muestreo, que también puede tener un impacto importante en la interpretación de los resultados. Hoy en día se recomienda añadir, cuando menos, intervalos de confianza.
- Los fenómenos adversos a menudo son fenómenos extremos, para los que las métricas habituales tienen poca o ninguna utilidad pues se trata de comportamientos asintóticos.
- La predecibilidad es dependiente del flujo y, por ende, la estratificación de los datos de verificación debería hacerse según ese flujo (patrones atmosféricos), superando la tradicional estratificación estacional o mensual.
- Las diferentes representaciones y escalas espaciales de modelos y observaciones son un elemento crítico de dificultas en verificación. Los métodos clásicos penalizan doblemente a los modelos de muy alta resolución. A raíz de este problema han surgido nuevas familias de métodos de verificación, los llamados métodos espaciales.

Si la verificación de modelos deterministas es un área delicada, podemos imaginar que la verificación de SPC lo es aún más. Comparar predicciones en forma de probabilidades con observaciones convencionales es complicado. Si las predicciones probabilistas se generan en forma de PDF ¿Qué propiedades son recomendables en una buena PDF? Hay dos grandes grupos de metodologías para dar respuesta a la pregunta.

Por un lado y, sobre todo, para evaluar la consistencia de las predicciones con las observaciones en el flujo a gran escala, se utilizan histogramas de rango y se contrasta la relación dispersión-error: la dispersión del SPC con respecto a su valor medio crece con el alcance predictivo y el error del valor medio con respecto a la observación también. Si crecen al mismo ritmo entonces mantienen una relación aproximadamente lineal y se dice que el SPC es estadísticamente consistente con las observaciones, fiable o calibrado.

Por otro lado y, en este otro gran grupo, para las variables meteorológicas de tiempo sensible, como la precipitación, el viento, la temperatura, etc., se mide la respuesta del sistema frente a eventos binarios (normalmente superación de umbrales, e. g. viento por encima de 70 km/h). Una serie de propiedades describen esa respuesta. La *agudeza* refleja que el sistema da probabilidades cercanas al 0 y al 1 y no probabilidades intermedias, menos útiles para el usuario. La *fiabilidad* es la consistencia entre las probabilidades previstas y las correspondientes frecuencias de observación condicionadas. La *resolución* es la mejora con respecto a la climatología muestral según los diferentes intervalos de probabilidad. Fiabilidad y resolución se representan en el llamado diagrama de fiabilidad o de atributos. La *discriminación* es la capacidad del sistema de discernir entre la ocurrencia y no ocurrencia del evento y se representa mediante las llamadas curvas ROC. Estas propiedades pueden construirse en un marco gráfico y, también, en marco teórico, partiendo de la generalización del error cuadrático medio al espacio de las probabilidades: el llamado BRIER *score*. Esta métrica puede descomponerse y las componentes tienen un significado muy preciso: fiabilidad, resolución e incertidumbre, en correspondencia con las propiedades en el marco gráfico. Todo ello conforma el enfoque llamado orientado a distribuciones, en particular la denominada *distribución conjunta* de predicciones y observaciones.

La predicción probabilista tiene valor si ayuda al usuario a tomar decisiones. Dentro del contexto de la respuesta frente a eventos binarios, el denominado *valor económico relativo* de un sistema predictivo nos indica, en términos económicos de ahorro con respecto a la climatología muestral, el valor del uso de nuestro sistema comparado con una predicción perfecta. Esta métrica de valor económico relativo permite comparar modelos deterministas y SPC, aunque no tiene por qué haber competencia entre ambos.

15.12 Enlaces de interés

En los siguientes enlaces se abordan métodos de verificación, pautas de coordinación internacional para la misma, así como información (relativamente) actualizada sobre calidad y valor de los modelos o SPC.

Página de los grupos de trabajo para verificación de la Organización Meteorológica Mundial (OMM), con un compendio completísimo sobre métodos de verificación (WWRP/WGNE Joint Working Group on Forecast Verification Research) (consultada 22-03-2018): http://www.cawcr.gov.au/projects/verification/

- Página de la NOAA, de HAROLD BROOKS, con recursos diversos relacionados con la verificación (NOAA Harold Brooks' *Weather Forecast Verification*, A collection of ideas, references, and resources for forecasters, researchers, teachers, and students) (consultada 22-03-2018): http://www.nssl.noaa.gov/users/brooks/public_html/verification/
- Módulo del programa COMET sobre *Uso inteligente de los productos derivados de los modelos*, que incluye apartados de verificación de los modelos (consultada 22-03-2018): https://www.meted.ucar.edu/training_module.php?id=797#.VXlnWqFCZwE
- Referencia de la OMM del programa de verificación, dentro del programa general de calidad (WMO > Programmes > AMP > PWS Home > Quality assurance > Verification) (consultada 22-03-2018): https://www.wmo.int/pages/prog/amp/pwsp/qualityassuranceverification_en.htm
- Pautas establecidas por la OMM para la verificación de predicciones de largo plazo (WMO - Standard Verification System (SVS) for Long-range Forecasts (LRF)) (consultada 22-03-2018): http://www.wmo.int/pages/prog/www/DPS/verification_systems.html
- Pautas establecidas por la OMM para la verificación de predicciones de largo plazo, un primer borrador (WMO - Standardised Verification System (SVS) for Long-Range Forecasts (LRF)) (consultada 22-03-2018): http://www.wmo.int/pages/prog/www/DPS/SVS-for-LRF.html
- Información actualizada de la calidad las predicciones del ECMWF (ECMWF > Forecasts > Quality of our forecasts > operational forecast evaluation > regularly updated on the ECMWF web site) (consultada 22-03-2018): https://www.ecmwf.int/en/forecasts/charts/catalogue/?f%5B0%5D=im_field_chart_type_2%3A606
- Información actualizada de la calidad de las predicciones del United States National Centers for Environmental Prediction (NCEP) (WPC's Model Diagnostics and Verification Page, NOAA > NCEP > NWS > Model Diagnostics) (consultada 22-03-2018): http://www.wpc.ncep.noaa.gov/html/model2.shtml

15.13 Referencias

[1] AHIJEVYCH, David y col. "Application of Spatial Verification Methods to Idealized and NWP-Gridded Precipitation Forecasts". En: *Weather and Forecasting* 24.6 (dic. de 2009), páginas 1485-1497. ISSN: 0882-8156. DOI: 10.1175/2009WAF2222298.1 (citado en páginas 218, 220).

[2] ANDERSON, J.L. L. "A Method for Producing and Evaluating Probabilistic Forecast from Ensemble Model Integration". En: *Journal of climate* 9.7 (1995), páginas 1518-1530. ISSN: 0894-8755. DOI: 10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2 (citado en página 222).

[3] BRADLEY, A. Allen y SCHWARTZ, Stuart S. "Summary Verification Measures and Their Interpretation for Ensemble Forecasts". En: *Monthly Weather Review* 139.9 (2011), páginas 3075-3089. ISSN: 0027-0644. DOI: 10.1175/2010MWR3305.1 (citado en página 218).

[4] BRADLEY, A. Allen, SCHWARTZ, Stuart S. y HASHINO, T. "Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score". En: *Weather and Forecasting* 23 (2008), página 992. DOI: 10.1175/2007WAF2007049.1 (citado en página 218).

[5] BUIZZA, Roberto. "Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system". En: *Monthly Weather Review* 125.1 (1997), páginas 99-119. ISSN: 0027-0644. DOI: 10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2 (citado en página 223).

[6] CALLADO, Alfons y col. "Ensemble Forecasting". En: *Climate Change and Regional/Local Responses*. Editado por RAY, Pallav. InTech, mayo de 2013. ISBN: 978-953-51-1132-0. DOI: 10.5772/55699 (citado en página 223).

[7] CANDILLE, Guillem y TALAGRAND, O. "Evaluation of probabilistic prediction systems for a scalar variable". En: *Quarterly Journal of the Royal Meteorological Society* 131.609 (2005), páginas 2131-2150. ISSN: 00359009. DOI: 10.1256/qj.04.71 (citado en páginas 208, 222, 231, 232).

[8] CANDILLE, Guillem y TALAGRAND, O. "Impact of observational error on the validation of ensemble prediction systems". En: *Quarterly Journal of the Royal Meteorological Society* 134.633 B (abr. de 2008), páginas 959-971. ISSN: 00359009. DOI: 10.1002/qj.268 (citado en página 218).

[9] CANDILLE, Guillem y col. "Verification of an Ensemble Prediction System against Observations". En: *Monthly Weather Review* 135.7 (2007), páginas 2688-2699. ISSN: 0027-0644. DOI: 10.1175/MWR3414.1 (citado en página 232).

[10] DAVIS, Christopher y col. "The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program". En: *Weather and Forecasting* 24.5 (oct. de 2009), páginas 1252-1267. ISSN: 0882-8156. DOI: 10.1175/2009WAF2222241.1 (citado en páginas 218, 220).

[11] EBERT, Elizabeth E. y GALLUS, William a. "Toward Better Understanding of the Contiguous Rain Area (CRA) Method for Spatial Forecast Verification". En: *Weather and Forecasting* 24.5 (oct. de 2009), páginas 1401-1415. ISSN: 0882-8156. DOI: 10.1175/2009WAF2222252.1 (citado en páginas 218, 220).

[12] EFRON, B. y TIBSHIRANI, R. "Improvements on cross-validation: The 632 plus bootstrap method". En: *Journal of the American Statistical Association* 92.438 (1997), página 548. ISSN: 0162-1459. DOI: 10.1080/01621459.1997.10474007 (citado en página 218).

[13] FERRANTI, L y CORTI, S. "Ensemble prediction skill in relation with large scale circulation patterns". En: *10th EMS Annual Meeting, 10th European Conference on Applications of Meteorology (ECAM) Abstracts,*

held Sept. 13-17, 2010 in Zürich, Switzerland. Sep. de 2010, página 769. URL: http://meetings.copernicus.org/ems2010 (citado en página 218).

[14] FERRO, Christopher a. T. "A Probability Model for Verifying Deterministic Forecasts of Extreme Events". En: *Weather and Forecasting* 22.5 (2007), páginas 1089-1100. ISSN: 0882-8156. DOI: 10.1175/WAF1036.1 (citado en página 218).

[15] GARCÍA-MOYA, José Antonio y col. "Predictability of short-range forecasting: A multimodel approach". En: *Tellus, Series A: Dynamic Meteorology and Oceanography* 63.3 (mayo de 2011), páginas 550-563. ISSN: 02806495. DOI: 10.1111/j.1600-0870.2010.00506.x (citado en páginas 221, 223).

[16] GHELLI, Anna y LALAURETTE, François. "Verifying precipitation forecasts using upscaled observations". En: *ECMWF Newsletter* 87 (2000), páginas 9-17 (citado en página 219).

[17] GHELLI, Anna y PRIMO, Cristina. "On the use of the extreme dependency score to investigate the performance of an NWP model for rare events". En: *Meteorological Applications* 16.4 (dic. de 2009), páginas 537-544. ISSN: 13504827. DOI: 10.1002/met.153 (citado en página 218).

[18] GILLELAND, Eric. "Confidence intervals for forecast verification". En: *NCAR Technical Note NCAR/TN-479+STR* (2010). DOI: 10.5065/D6WD3XJM (citado en página 218).

[19] GILLELAND, Eric y col. "Intercomparison of Spatial Forecast Verification Methods". EN. En: *Weather and Forecasting* 24.5 (oct. de 2009), páginas 1416-1430. ISSN: 0882-8156. DOI: 10.1175/2009WAF2222269.1 (citado en páginas 218, 220).

[20] GUTIÉRREZ, J.˜M. y col. "Clustering Methods for Statistical Downscaling in Short-Range Weather Forecasts". En: *Monthly Weather Review* 132.1997 (2004), página 2169. ISSN: 0027-0644. DOI: 10.1175/1520-0493(2004)132<2169:CMFSDI>2.0.CO;2 (citado en página 231).

[21] HAMILL, Thomas M. "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". En: *Monthly Weather Review* 129.3 (2001), páginas 550-560. ISSN: 0027-0644. DOI: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2 (citado en página 222).

[22] HAMILL, Thomas M. y COLUCCI, Stephen J. "Verification of Eta-RSM short-range ensemble forecasts". En: *Monthly Weather Review* 125.6 (1997), páginas 1312-1327 (citado en página 222).

[23] HAMILL, Thomas M. y COLUCCI, Stephen J. "Evaluation of Eta-RSM Ensemble Probabilistic Precipitation Forecasts". En: *Monthly Weather Review* 126.3 (1998), páginas 711-724. ISSN: 0027-0644. DOI: 10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2 (citado en página 222).

[24] HANLEY, James A y MCNEIL, Barbara J. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." En: *Radiology* 143.1 (1982), páginas 29-36 (citado en página 231).

[25] HERSBACH, Hans. "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems". En: *Weather and Forecasting* 15.5 (oct. de 2000), páginas 559-570. ISSN: 0882-8156. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2 (citado en página 232).

[26] JOLLIFFE, Ian T. y STEPHENSON, David B. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2003, página 254. ISBN: 0470864419. DOI: 10.1016/j.ijforecast.2005.11.002 (citado en páginas 208, 218, 224, 227, 231, 232).

[27] JOLLIFFE, Ian T. y STEPHENSON, David B. "Comments on "Discussion of Verification Concepts in Forecast Verification: A Practitioner's Guide in Atmospheric Science ". En: *Weather and Forecasting* 20.5 (2005), páginas 796-800. ISSN: 0882-8156. DOI: 10.

1175/WAF877.1 (citado en páginas 218, 231).

[28] JOLLIFFE, Ian T. y STEPHENSON, David B. "Introduction". En: *Forecast Verification*. John Wiley & Sons, Ltd, 2011, páginas 1-9. ISBN: 9781119960003. DOI: 10.1002/9781119960003.ch1 (citado en páginas 210, 218, 224, 231, 232).

[29] KHARIN, Viatcheslav V. y ZWIERS, Francis W. "On the ROC score of probability forecasts". En: *Journal of Climate* 16.24 (2003), páginas 4145-4150. ISSN: 08948755. DOI: 10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2 (citado en página 231).

[30] LALAURETTE, François. "Early detection of abnormal weather conditions using a probabilistic extreme forecast index". En: *Quarterly Journal of the Royal Meteorological Society* 129.594 (oct. de 2003), páginas 3037-3057. ISSN: 00359009. DOI: 10.1256/qj.02.152 (citado en página 218).

[31] LAVERS, David A. y col. "ECMWF Extreme Forecast Index for water vapor transport: A forecast tool for atmospheric rivers and extreme precipitation". En: *Geophysical Research Letters* 43.22 (nov. de 2016), páginas 11, 811-852, 858. ISSN: 00948276. DOI: 10.1002/2016GL071320 (citado en página 218).

[32] LEITH, C E. "Theoretical skill of Monte Carlo forecasts". En: *Monthly Weather Review* 102.6 (1974), páginas 409-418 (citado en página 221).

[33] MURPHY, Allan H. "Forecast verification: Its Complexity and Dimensionality". En: *Monthly Weather Review* 119.7 (1991), páginas 1590-1601. ISSN: 0027-0644. DOI: 10.1175/1520-0493(1991)119<1590:FVICAD>2.0.CO;2 (citado en páginas 210, 224).

[34] MURPHY, Allan H. *Climatology, persistence, and their linear combination as standards of reference in skill scores*. 1992. DOI: 10.1175/1520-0434(1992)007<0692:CPATLC>2.0.CO;2 (citado en página 210).

[35] MURPHY, Allan H. "What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting". En: *Weather and Forecasting* 8.2 (jun. de 1993), páginas 281-293. ISSN: 0882-8156. DOI: 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2 (citado en página 208).

[36] MURPHY, James M. "The impact of ensemble forecasts on predictability". En: *Quarterly Journal of the Royal Meteorological Society* 114.480 (1988), páginas 463-493. DOI: 10.1002/qj.49711448010 (citado en página 221).

[37] PAIMAZUMDER, Debasish y MÖLDERS, Nicole. "Theoretical assessment of uncertainty in regional averages due to network density and design". En: *Journal of Applied Meteorology and Climatology* 48.8 (ago. de 2009), páginas 1643-1666. ISSN: 15588424. DOI: 10.1175/2009JAMC2022.1 (citado en página 219).

[38] PERSSON, Anders. *Verification of Probability Ensemble forecasts*. Informe técnico. 2006 (citado en página 221).

[39] PERSSON, Anders. "User guide to ECMWF forecast products". En: *Ecmwf* March (2011), página 127 (citado en página 221).

[40] PRIMO, Cristina y GHELLI, Anna. "The affect of the base rate on the extreme dependency score". En: *Meteorological Applications* 16.4 (dic. de 2009), páginas 533-535. ISSN: 13504827. DOI: 10.1002/met.152 (citado en página 218).

[41] RODWELL, Mark J. y col. "A new equitable score suitable for verifying precipitation in numerical weather prediction". En: *Quarterly Journal of the Royal Meteorological Society* 136.650 (2010), páginas 1344-1363. ISSN: 00359009. DOI: 10.1002/qj.656 (citado en página 219).

[42] SANTOS, Carlos y GHELLI, Anna. "Observational probability method to assess ensemble precipitation forecasts". En: *Quarterly Journal of the Royal Meteorological Society* 138.662 (ene. de 2012), páginas 209-221.

ISSN: 00359009. DOI: 10.1002/qj.895 (citado en página 218).

[43] STENSRUD, David J. y YUSSOUF, Nusrat. "Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system". En: *Weather and Forecasting* 22.1 (2007), páginas 3-17 (citado en página 208).

[44] STEPHENSON, David B. y col. "The extreme dependency score: a non-vanishing measure for forecasts of rare events". En: *Meteorological Applications* 15.1 (mar. de 2008), páginas 41-50. ISSN: 13504827. DOI: 10.1002/met.53 (citado en página 218).

[45] WEISHEIMER, A y col. "ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions - Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs". En: *Geophysical research letters* 36.21 (2009). DOI: 10.1029/2009GL040896 (citado en páginas 227, 228).

[46] WERNLI, Heini y col. "SAL-A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts". En: *Monthly Weather Review* 136.11 (nov. de 2008), páginas 4470-4487. ISSN: 0027-0644. DOI: 10.1175/2008MWR2415.1 (citado en páginas 218, 220).

[47] WHITAKER, Jeffrey S. y LOUGHE, Andrew F. "The Relationship between Ensemble Spread and Ensemble Mean Skill". En: *Monthly Weather Review* 126.12 (1998), páginas 3292-3302. ISSN: 0027-0644. DOI: 10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2 (citado en páginas 221, 223).

[48] WILKS, Daniel S. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 2011, página 676. ISBN: 9780123850225. URL: https://www.sciencedirect.com/bookseries/international-geophysics/vol/100 (citado en páginas 208, 210, 218, 224, 227, 232).

[49] ZHANG, Fuqing. "Dynamics and Structure of Mesoscale Error Covariance of a Winter Cyclone Estimated through Short-Range Ensemble Forecasts". En: *Monthly Weather Review* 133.10 (oct. de 2005), páginas 2876-2893. ISSN: 0027-0644. DOI: 10.1175/MWR3009.1 (citado en página 218).

[50] ZIEHMANN, Christine. "Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models". En: *Tellus, Series A: Dynamic Meteorology and Oceanography* 52.3 (mayo de 2000), páginas 280-299. ISSN: 02806495. DOI: 10.3402/tellusa.v52i3.12266 (citado en página 221).

[51] ZSOTER, Ervin, PAPPENBERGER, Florian y RICHARDSON, David. "Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index". En: *Meteorological Applications* 22.2 (abr. de 2015), páginas 236-247. ISSN: 13504827. DOI: 10.1002/met.1447 (citado en página 218).