



Nuevas técnicas de agrupamiento de predicciones por conjuntos aplicadas a campos de superficie

Autor: Antonio Rodríguez Martínez

Institución: Agencia Estatal de Meteorología (Aemet)

Resumen

Los errores en las predicciones obtenidas a partir de modelos numéricos de predicción del tiempo deterministas son debidos a una combinación de los errores del análisis de partida y a deficiencias en la formulación del modelo. La estrategia seguida para superar estas limitaciones ha sido desarrollar una aproximación probabilística a la predicción numérica del tiempo a partir de los Sistemas de Predicción por Conjuntos o Ensemble Prediction System (EPS). Con diferentes métodos se computa un conjunto de predicciones válidas para un determinado rango temporal y para la misma zona geográfica denominadas miembros del EPS. Se dispone así de un sistema muy útil que cuantifica la incertidumbre de la predicción a través de la dispersión, muestra escenarios alternativos al modelo determinista y permite generar predicciones probabilísticas.

En AEMET se está ensayando en la actualidad el nuevo software de 'clustering' (agrupación óptima de miembros) del Centro Europeo de Predicción a Medio Plazo (ECMWF) que permite sintetizar la información contenida en los 51 miembros de su EPS. Se presentan en este documento una descripción general de la metodología utilizada por esta técnica de agrupamiento así como los resultados obtenidos de su aplicación sobre una región que incluye la península ibérica.

El agrupamiento, realizado diariamente para las pasadas de las 00 UTC y 12 UTC del modelo del ECMWF, se ha basado en la similitud entre los campos próximos a superficie sobre un dominio de 25°x35° para diferentes ventanas temporales utilizándose una resolución de grid de 0.25°x0.25°. Los resultados del agrupamiento, que proporciona un número variable de escenarios meteorológicos (de 1 a 6) alternativos y representativos del conjunto, se han aplicado a parámetros de superficie como el viento a 10 metros, presión a nivel del mar y temperatura a 2 metros.

Palabras clave: predicción meteorológica; predicción numérica; predicción probabilística; predicción por conjuntos; clustering; técnicas de agrupamiento; EPS

1. Introducción

Los errores en las condiciones iniciales, no importa como sean de pequeños, imponen un límite en la validez temporal de la predicción debido a que la atmósfera es un sistema caótico, donde estados que parecen idénticos, pero que en realidad no lo son, pueden evolucionar hacia estados claramente diferentes.

Los errores en las predicciones obtenidas a partir de modelos numéricos de predicción del tiempo deterministas son debidos a una combinación de los errores del análisis de partida y a deficiencias en la formulación del modelo.

La estrategia seguida por diversos Servicios Meteorológicos Nacionales para superar las limitaciones de los modelos deterministas ha sido desarrollar Sistemas de Predicción por Conjuntos o Ensemble Prediction System (EPS). Para profundizar en este tema se puede consultar la publicación “Ensemble Forecasting” (Callado et al., 2013).

En este documento nos referiremos al EPS desarrollado Centro Europeo de Predicción a Medio Plazo (ECMWF) basado en la integración de un conjunto de 50 análisis perturbados, más uno sin perturbar llamado Control, mediante un modelo numérico de menor resolución que su modelo operativo determinista, obteniéndose así un conjunto o EPS de 51 miembros.

El fundamento del EPS radica en considerar los errores del análisis de partida como una de las causas principales de los errores de la predicción. Así el análisis de Control se perturba combinando diversas técnicas para intentar representar los errores que éste contiene. Para perturbar las condiciones iniciales se utiliza una combinación de la técnica de vectores singulares y perturbaciones producidas por conjuntos de asimilaciones de datos y para representar las incertidumbres en la formulación del modelo se utilizan técnicas de perturbación estocásticas. Más detalles se pueden encontrar en el documento “User guide to ECMWF forecast products” (Persson & Andersson, 2013).

El EPS es un sistema de predicción muy útil ya que cuantifica la incertidumbre de la predicción a través de la dispersión, muestra escenarios alternativos al modelo operativo y a partir de él se pueden generar predicciones probabilísticas. Existe un amplio rango de productos derivados del EPS, a saber:

1.1. Sellos

Los sellos del ECMWF muestran individualmente la predicción determinista de alta resolución, 16Km, así como la predicción de Control y la de los 50 miembros de su EPS a una resolución de 32 Km, para la presión a nivel del mar y la temperatura a 850 hPa (ver Figura 1). Los mapas se presentan como referencia, por ejemplo, para explicar sinópticamente la dispersión de las predicciones, siendo difícil intentar determinar un “Miembro del Día”.

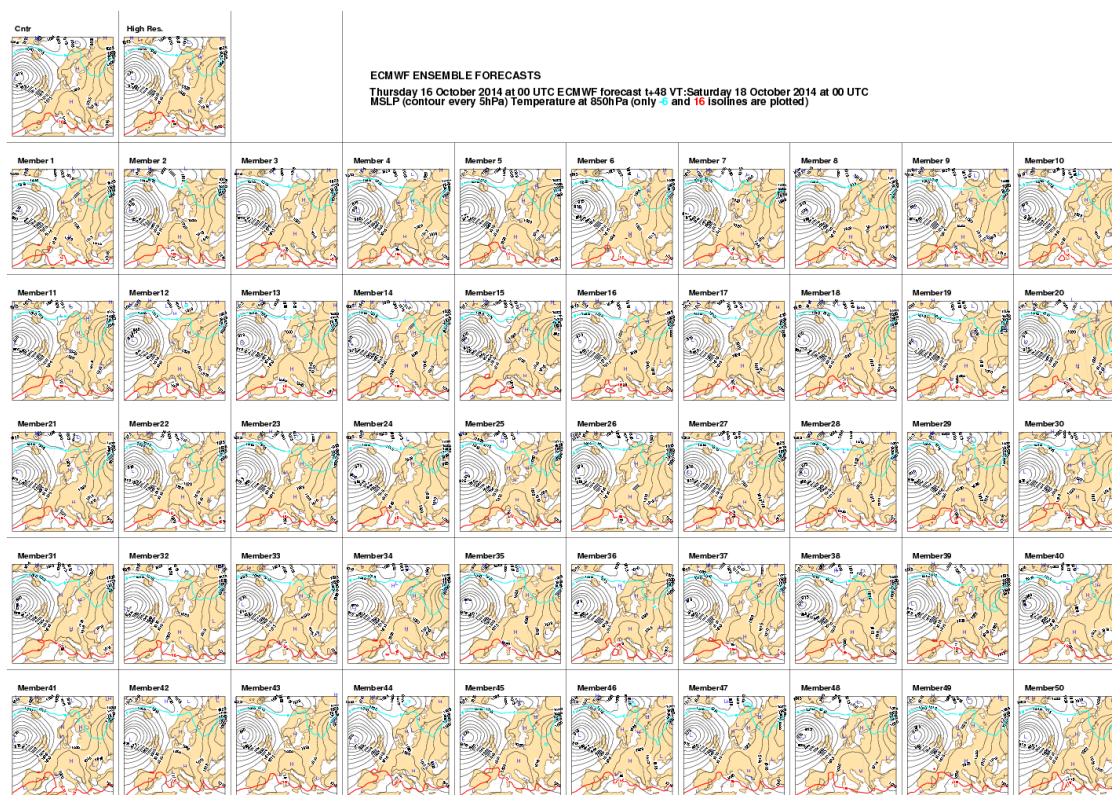


Figura 1. Sellos correspondientes a la predicción a dos días, basada en el EPS del ECMWF del 18 de octubre de 2014 a las 00 UTC para la presión a nivel del mar (hPa) y la temperatura a 850 hPa (isolíneas de -6 y 16 °C). Cortesía del ECMWF(2014).

1.2. Meteogramas

Los meteogramas o EPS-gramas ofrecen una representación probabilística de las predicciones del EPS a 10 o 15 días de ciertos parámetros para un lugar determinado. La Figura 2 muestra la evolución en 10 días de la cobertura nubosa, la precipitación acumulada en 6 horas, el viento a 10 m y la temperatura a 2m. En azul se representa la predicción del modelo determinista de alta resolución, y en rojo la predicción de Control. Los box-plot representan la distribución de probabilidad de los 50 miembros perturbados incluyendo los extremos, la mediana y los percentiles 10, 25, 50, 75 y 90.

ENS Meteogram

San Salvador, El Salvador 13.63°N 89.16°W (EPS land point) 699 m
High Resolution Forecast and ENS Distribution Monday 6 October 2014 12 UTC

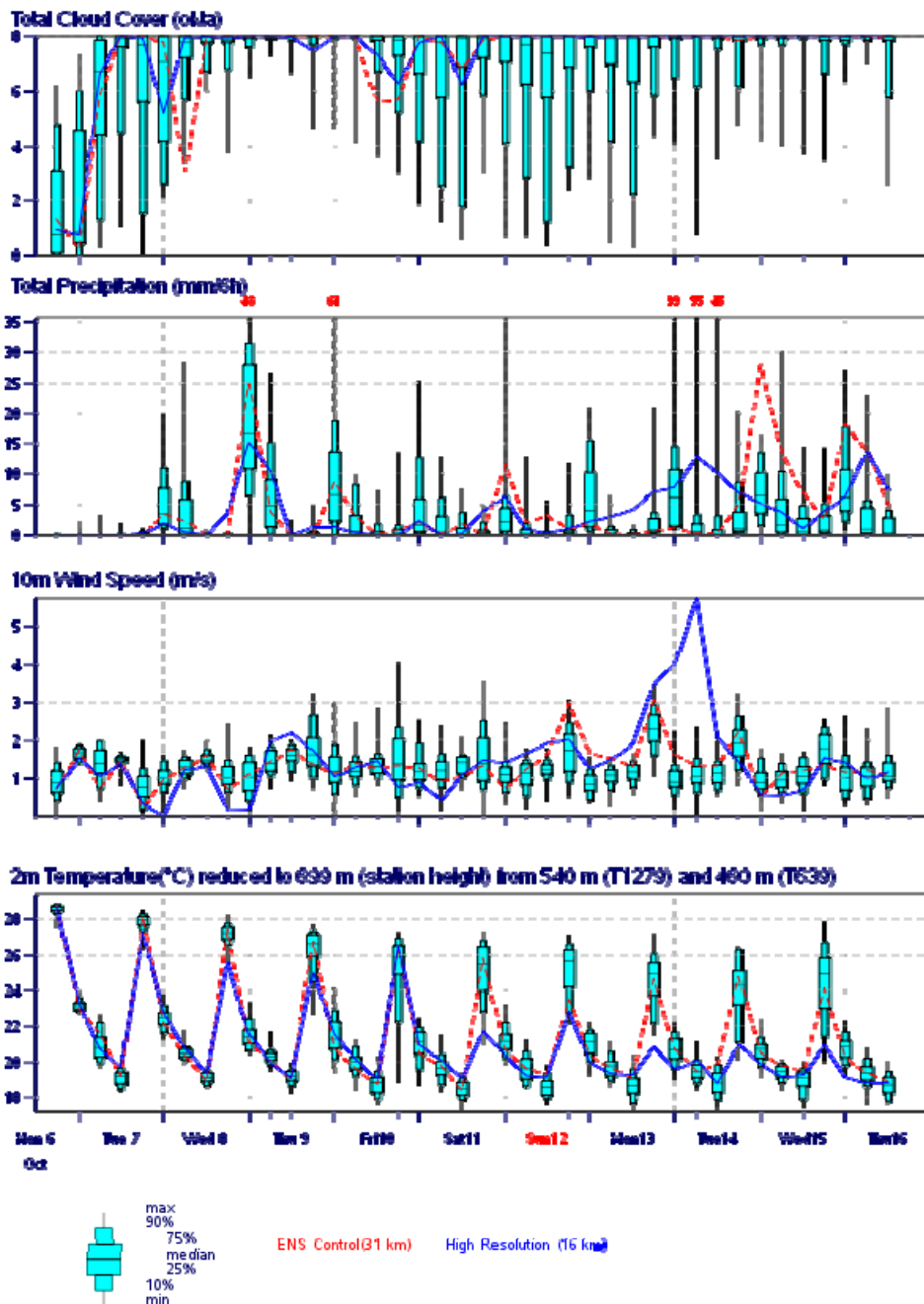


Figura 2. Meteogramas para la predicción a 10 días, basada en el EPS del ECMWF del 6 de octubre de 2014 a las 12 UTC. La línea continua azul corresponde al modelo determinista de alta resolución y la discontinua roja a la predicción de Control. Cortesía del ECMWF(2014).

1.3. Mapas de probabilidad

A partir de los miembros del EPS del ECMWF se generan en AEMET mapas de probabilidad según umbrales, para diferentes variables meteorológicas y diferentes rangos de predicción. La Figura 3 representa diferentes intervalos de probabilidad de precipitación acumulada en 24 horas, superior a 1, 5, 10 y 20 mm, para un alcance entre 3 y 5 días.

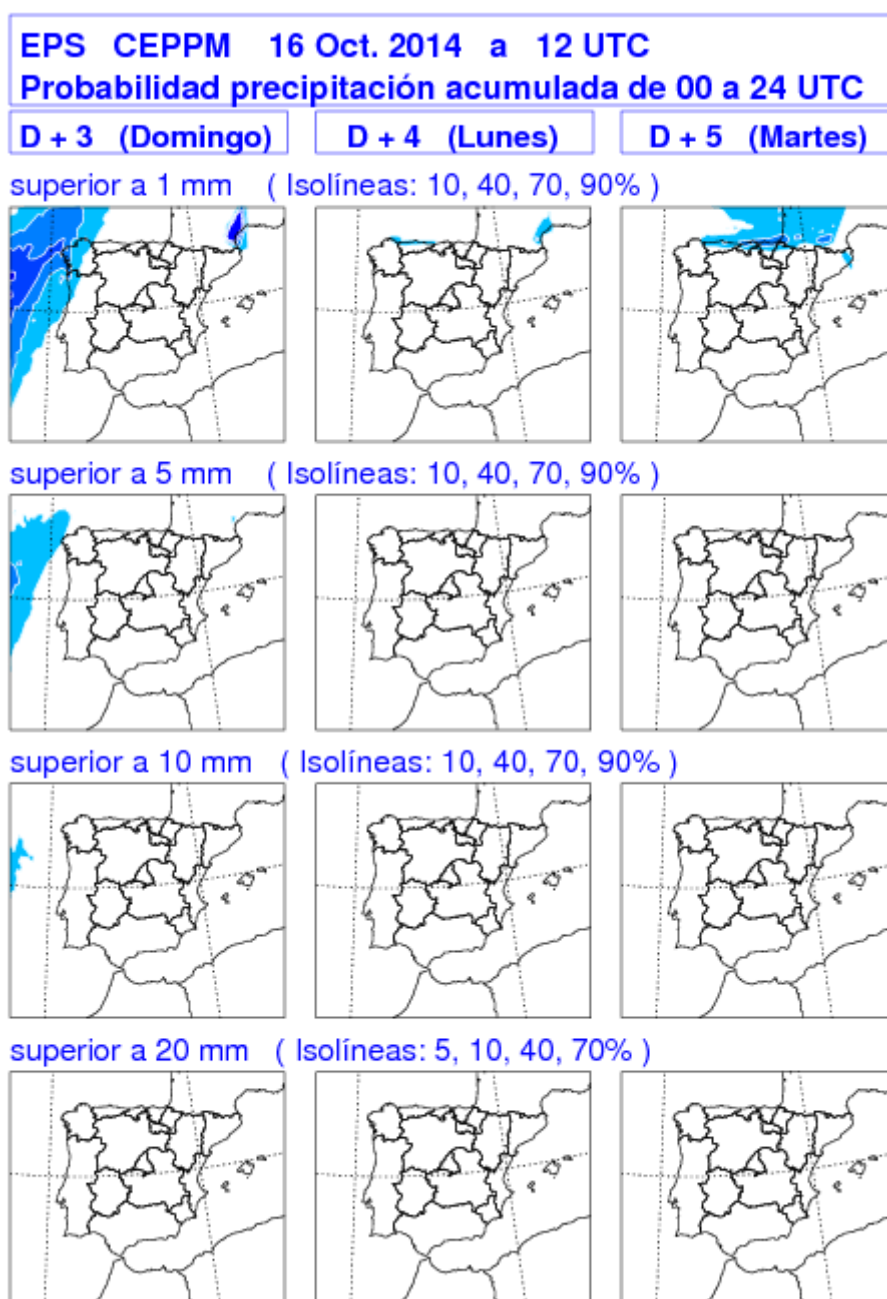


Figura 3. Probabilidad de precipitación acumulada de 00 a 24 UTC superior a 1, 5, 10 y 20 mm, basada en EPS del ECMWF del 16 de octubre de 2014 a las 12 UTC.

Existen otros productos derivados del EPS. Los espaguetis representan la 51 predicciones en el mismo mapa, pero solamente una isolínea de forma que se pueda observar, a grandes rasgos, el comportamiento de todos los miembros sobre una zona determinada. Los mapas de promedio y dispersión permiten juzgar el alcance en el tiempo del valor de la información del promedio del EPS en relación al flujo a gran escala, así como la dispersión asociada a la incertidumbre en la predicción de ciertas variables. También se generan índices para tiempo severo (Extrem Forcast Index-EFI) y mapas y diagramas para el seguimiento de ciclones tropicales y extra-tropicales. Los 51 miembros del EPS se pueden agrupar para formar los denominados “clusters” o grupos, de estos productos nos ocuparemos en los siguientes apartados.

2. Nueva técnica de “Clustering” en el ECMWF

Para sintetizar la gran cantidad de información producida por el EPS y visualizar los diferentes escenarios que contiene es necesario realizar un agrupamiento de aquellos miembros que son más similares entre sí, según una cierta norma. Para cada grupo se calcula la media de todos los miembros del EPS que incluye y se identifica su miembro más representativo. Se obtienen así nuevas “predicciones” que facilitan el manejo de la información y resaltan las partes predecibles.

El nuevo software del ECMWF (operacional desde Noviembre 2010) utiliza dos tipos de agrupamiento, un agrupamiento diario basado en escenarios meteorológicos y un segundo agrupamiento basado en escenarios climatológicos.

El agrupamiento diario se basa en la similaridad en la evolución del campo de geopotencial de 500 hPa sobre una región que cubre el Atlántico Norte y Europa (30°N-75°N, 20°W-40°E). Se obtienen así un número variable de escenarios meteorológicos (de 1 a 6) definidos cada uno de ellos por el miembro más próximo al centroide o “centro de gravedad” de cada grupo, representando el rango de patrones de flujo sinóptico contenidos en el EPS.

En una segunda etapa cada escenario del EPS se asigna al régimen más cercano de un conjunto de uno de cuatro regímenes climatológicos fijos y estacionales (estación cálida y fría) definidos en términos de flujo a gran escala (Oscilación del Atlántico Norte positiva y negativa, Dorsal Atlántica y Situación de Bloqueo sobre Europa y el Atlántico). Esto permite tener una medida objetiva de las diferencias entre los escenarios en términos del flujo a gran escala y un cálculo de medidas de skill o habilidad de las predicciones dependientes del flujo.

En AEMET se está ensayando el nuevo software de agrupamiento del ECMWF en base a parámetros de superficie por lo que solamente se realiza en agrupamiento diario no pudiéndose determinar la asignación de régimen climatológico, definido únicamente en términos de la evolución del geopotencial de 500 mb sobre el Atlántico Norte y Europa y no calculado por el software distribuido por el ECMWF. La información completa de ambos tipos de agrupamiento se puede consultar en el documento “New clustering products” (Ferranti & Corti, 2011).

3. Nueva técnica de “Clustering” en AEMET

3.1. Agrupamiento diario del EPS en escenarios para parámetros de superficie.

En AEMET se está utilizando el software del ECMWF realizándose en la actualidad un agrupamiento diario basado en el campo de presión a nivel del mar en lugar del geopotencial de 500 hPa debido a un interés en la obtención de agrupamientos de parámetros de superficie. El agrupamiento se realiza para las pasadas de las 00 UTC y 12 UTC sobre un dominio que incluye la península ibérica y parte del Atlántico (30°N-55°N, 25°W-10°E), utilizándose una resolución de grid de 0.25°x0.25°. El resultado se aplica posteriormente al viento a diez metros y a la temperatura a dos metros.

El agrupamiento diario se puede realizar para diferentes ventanas temporales elegidas: 3-4 días, 5-7 días, etc.. Para la ventana de 3 a 4 días se presentan los escenarios para H+72, H+84 y H+96 (a 72, 84 y 96 horas respectivamente) mostrando la evolución del miembro representante del grupo desde H+72 a H+96. Para la ventana de 5 a 7 días se presentan los escenarios para H+120, H+144 y H+168 mostrando la evolución del miembro representante del grupo desde H+120 a H+168, y así para cada ventana elegida para hacer el agrupamiento siempre y cuando éstas no sean superiores a cinco días.

Cada grupo se describe por su centroide y no por la media de sus miembros. Para preservar la continuidad sinóptica el representante de cada grupo debe ser el mismo miembro en cada ventana temporal. Para representar cada grupo se utilizará no la media sino el miembro del grupo más cercano al centroide del mismo utilizándose como medida de distancia la raíz cuadrada de las diferencias cuadráticas medias.

El representante de cada grupo no debe considerarse como un sustituto de la media ni utilizarse como una predicción determinista. Los miembros representativos de cada grupo serán referidos como “escenarios del EPS” y serán identificados por su número de perturbación dentro del conjunto de predicciones, etiquetándose la predicción de Control como miembro 0. La “población” o número de miembros en cada cluster, considerados a priori todos igualmente probables, será un indicativo de la probabilidad o peso de cada escenario.

El número de grupos dependerá de la distribución del conjunto de predicciones. Si la distribución incluye pocos y bien diferenciados grupos de predicciones similares (distribución multimodal) los diferentes escenarios representaran óptimamente las posibles condiciones meteorológicas.

Así en la Figura 4 se representan los escenarios de la presión a nivel del mar en hPa obtenidos a partir del agrupamiento del 13 de octubre de 2014 a las 12 UTC para un rango de predicción entre HH+72 y H+96. El grupo 1 incluiría 24 miembros de los 51 habiéndose identificado el miembro 22 como representativo del cluster. Análogamente el grupo 2 incluye 17 predicciones y estaría representado por el miembro 8 del EPS. El grupo 3 sería el menos probable incluyendo 10 miembros y siendo representado por la perturbación 26.

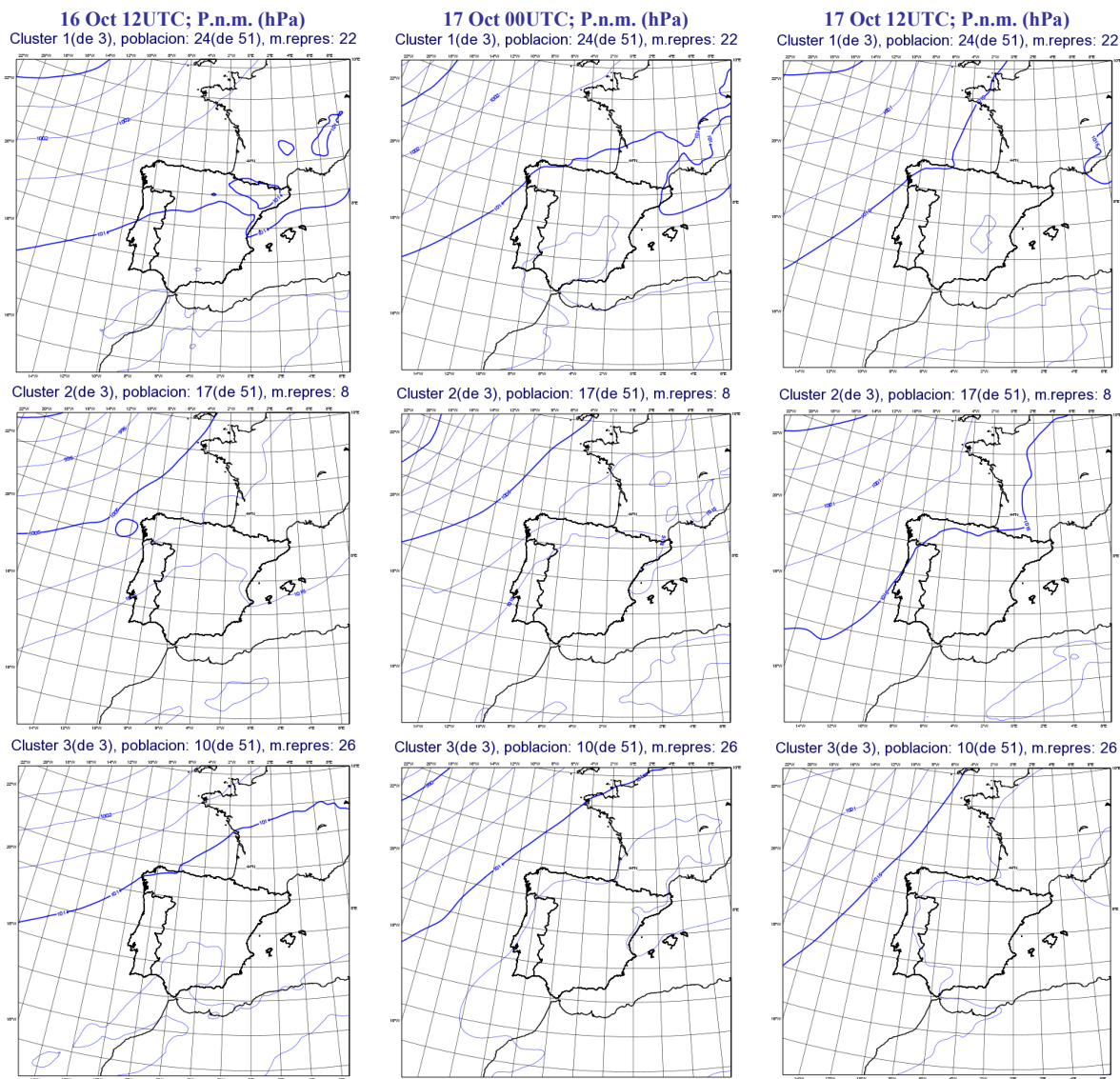


Figura 4. Escenarios para la ventana temporal de 72-96 horas basados en el EPS del 13 de octubre 2014 a las 12 UTC para la presión a nivel del mar en hPa.

En la Figura 5 se muestran los escenarios proporcionados por el mismo agrupamiento descrito en el ejemplo anterior aplicado esta vez al viento a 10 metros en nudos representados en una región más pequeña para su mejor visualización.

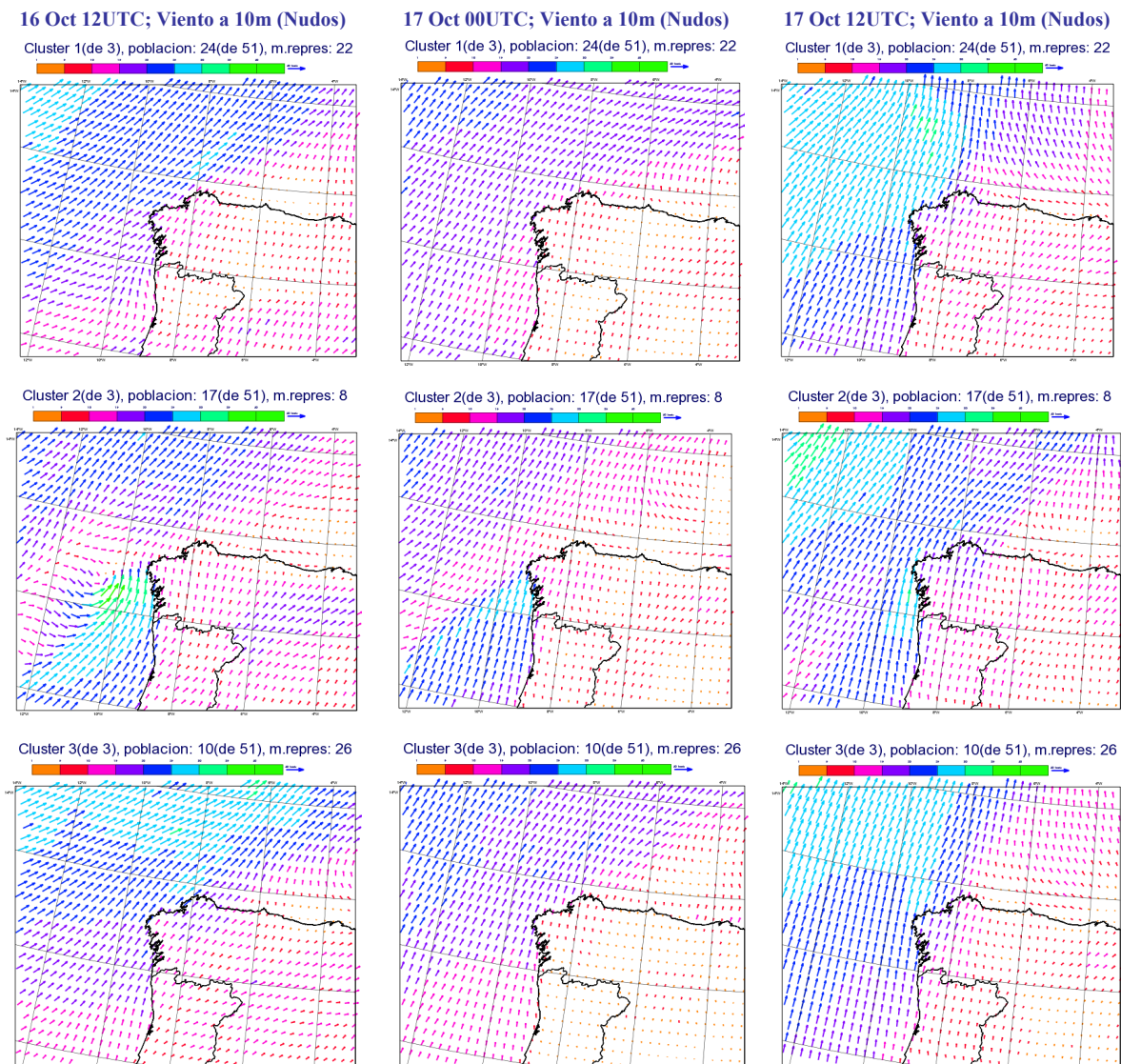


Figura 5. Escenarios para la ventana temporal de 72-96 horas basados en el EPS del 13 de octubre de 2014 a las 12 UTC para el viento a 10 metros en nudos.

Si la distribución del EPS es muy homogénea y el algoritmo no conduce a la formación de grupos claramente separados (distribución unimodal), todos los miembros estarán contenidos en un solo grupo que será representado por el miembro más próximo a la media. Así en la Figura 6 se muestra el único grupo obtenido a partir del algoritmo para el día 28 de septiembre de 2014 a las 00 UTC para la presión a nivel del mar, el viento a 10 metros y la temperatura a 2 metros.

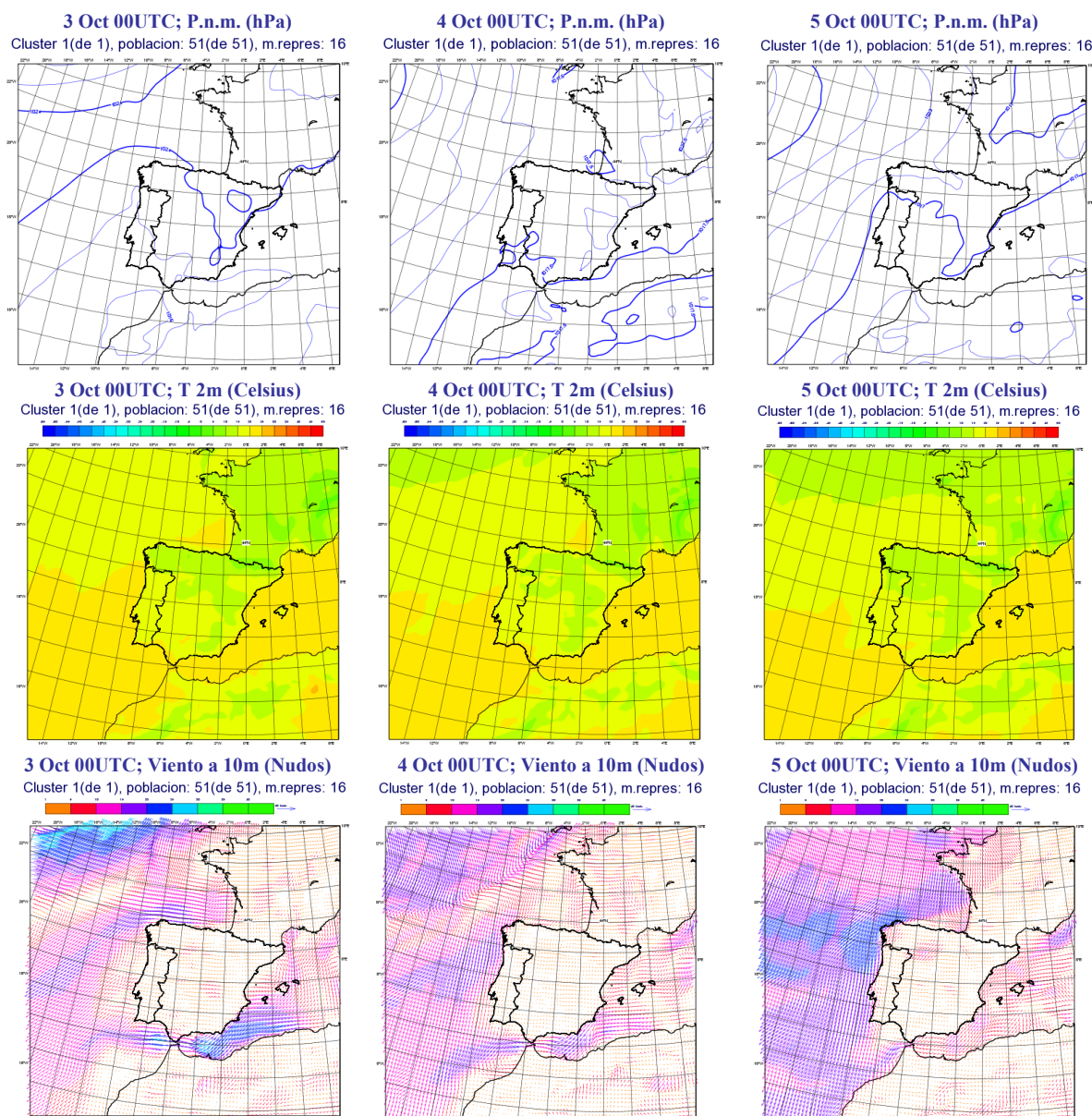


Figura 6. Escenarios para la ventana temporal de 120-168 horas basados en el EPS del 28 de septiembre de 2014 a las 00 UTC para la presión a nivel del mar (hPa), temperatura a 2 m (°C) y el viento a 10 metros (nudos).

La existencia de una gran dispersión (desviación estándar de la distribución del conjunto de predicciones) no conduce necesariamente a la generación de un grupo y por otra parte, incluso en el caso de que la dispersión sea pequeña, se formarán grupos si la partición es posible. Una gran dispersión del EPS no conduce automáticamente a más grupos ni una dispersión pequeña conduce necesariamente a menos.

Dado que se pretende sintetizar toda la información del EPS, el máximo número de grupos permitidos se limita a seis, y debido a que el agrupamiento se hace

independientemente para cada ventana temporal es muy probable que se generen más grupos en el corto que en el largo alcance.

3.2. Metodología para identificar los escenarios del EPS

El Nuevo agrupamiento ECMWF se basa en método K-medias, algoritmo particional que permite reasignar los miembros a diferentes grupos durante el proceso de agrupamiento hasta conseguir una partición estable y óptima. Esta metodología se puede resumir en cuatro etapas:

3.2.1. Identificación de un espacio de fases adecuado para la obtención del agrupamiento.

Las técnicas de clustering son efectivas únicamente si se aplican en un espacio de fases L-dimensional con $L < N$, siendo N el número de elementos en el dataset, en nuestro caso $N=51$ que es el número de miembros en el EPS. Debido a que el área en que se computa el clustering contiene muchos más de 51 puntos de grid es necesario en primer lugar transformar las predicciones a un nuevo sistema de coordenadas de mucha menor dimensión utilizando la denominada técnica de análisis de componentes principales que reduce la dimensión del espacio preservando el máximo de varianza de la muestra.

Para ello la base del nuevo espacio se forma con aquellos vectores donde la muestra proyectada presenta mayor varianza. Los vectores de esta base, denominados Funciones Ortogonales Empíricas (EOF's) son de enorme utilidad en meteorología pues son los patrones dominantes en el sentido de la variabilidad que representan en la muestra. Para cada ventana temporal y campo elegido, la presión a nivel del mar en nuestro caso, se realiza una descomposición en EOF's calculándose los grupos en un espacio de fases reducido definido por las L primeras EOF's principales de modo que la varianza acumulada explicada sea igual o superior al 80% de la varianza del EPS total.

Las componentes principales (CPs), proyección de los miembros del EPS sobre las EOFs, proporcionan las coordenadas de las predicciones en el nuevo espacio que serán utilizadas para el cómputo de los grupos. Los grupos se calculan para una determinada ventana temporal por lo que se utilizan EOFs extendidas en el tiempo, concretamente las calculadas para el rango de predicción de 5 a 7 días (HH+120 a HH+168).

3.2.2. Cálculo de la partición óptima del conjunto de datos

Para un número K asignado de antemano se busca la partición óptima del conjunto de datos en K grupos. Para ello se toman al azar K grupos iniciales. Se calculan las distancias de los miembros a los centroides de los grupos y se reasignan a los que estén más próximos. Se vuelven a recalcular los centroides de los K grupos después de las reasignaciones repitiéndose los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los miembros se estabilicen en algún grupo. El centroide de un grupo se define como la media de las coordenadas de todos los estados incluidos en dicho grupo.

El algoritmo se ejecuta muchas veces partiendo de diferentes K grupos iniciales calculándose en cada agrupamiento el cociente r_k^* de la varianza entre los centroides (ponderados por su población) y la media de las varianzas dentro de cada grupo. El agrupamiento que maximiza este cociente es el elegido.

El objetivo que se persigue al formar los grupos es que los centroides estén lo más separados entre sí como sea posible y que los miembros dentro de cada grupo estén muy próximos al centroide.

3.2.3 Estimación de la significación de una partición K dada

El objetivo es comparar la fuerza del agrupamiento con la esperada a partir de una función de distribución de referencia apropiada (distribución Gaussiana multidimensional). Para decidir si la hipótesis nula de multinormalidad se rechaza se generan un gran número M de muestras sintéticas con ciertas medidas estadísticas similares a la muestra original a partir de simulaciones de Montecarlo y procesos de Markov.

Para cada conjunto de datos simulados se realiza un análisis de agrupamiento, registrándose para cada partición K el cociente r_{mk} entre la varianza entre los centroides de cada grupo y la media de las varianzas dentro de cada grupo.

Como los datos sintéticos se asume siguen una distribución unimodal, la proporción P_k de muestras sintéticas con $r_{mk} < r_k^*$ será una medida de la significación de la partición K de los datos reales, y $1-P_k$ será el correspondiente nivel de confianza para la existencia de K grupos.

3.2.4 Elección del número de grupos más adecuado.

El método K -medias precisa de la especificación previa del número de grupos de la partición. En el agrupamiento diario la partición con la mayor significación estadística respecto a las distribuciones multi-normales predefinidas determina el criterio de elección del número óptimo de grupos.

El número de grupos se determina de la siguiente manera:

- Si la significación de todas las particiones consideradas, de 2 a 6 grupos, es inferior a un umbral mínimo de 55% no se formarán grupos.
- Si se alcanza el umbral mínimo, la partición con mayor significación será la elegida.
- Si más de una partición supera la significación del 95%, aquella con menor número de grupos será la elegida.

Se descartarán además particiones con una varianza interna media de sus grupos inferior al 50% de la varianza media climatológica del EPS, calculada como media del "spread" o dispersión del EPS los 30 días previos con el objetivo de limitar así la ocurrencia de particiones con un número grande de grupos (limitado a seis por el algoritmo). Teniendo en cuenta que el spread del EPS es función del alcance de la predicción, se añade cierta consistencia en la población de los grupos entre las diferentes ventanas temporales para las que se realiza el agrupamiento.

La información completa esta técnica utilizada por el nuevo algoritmo de agrupamiento del ECMWF se encuentra el documento “New clustering products” (Ferranti & Corti, 2011).

4. Resumen y trabajo futuro

El nuevo software de clustering del ECMWF, diseñado para calcular los grupos para un solo parámetro, fecha y tiempo de pasada de su modelo determinista, ha sido adaptado en AEMET para calcular los grupos para varios parámetros de superficie en una sola ejecución. Dicho software se ejecuta en el ECMWF, controlado desde el área de Innovación de AEMET, utilizando los 51 miembros del EPS para las pasadas de las 00 y 12 UTC.

El agrupamiento se realiza en base a la evolución de la presión a nivel del mar sobre un dominio de 25°x35° que cubre la península ibérica y una zona suficientemente amplia del Atlántico para incluir las perturbaciones que afectarían al comportamiento de las diferentes variables meteorológicas en nuestra región. El agrupamiento se realiza a una resolución de grid de 0.25°x0.25°.

En un futuro se realizará el agrupamiento en base a otros parámetros de superficie utilizando dominios que abarquen otras regiones (actualmente no se cubre Canarias). El agrupamiento se aplicará a otros parámetros de superficie además de la presión a nivel del mar, viento a 10 metros y temperatura a 2 metros, como son la precipitación acumulada en diferentes intervalos temporales, la racha máxima de viento, etc. Un trabajo de verificación frente a las variables meteorológicas observadas también es necesario para elegir el parámetro mas adecuado para realizar el agrupamiento, el dominio más apropiado, y cuantificar la calidad del pronóstico de las diferentes variables meteorológicas utilizadas en el agrupamiento.

5. Referencias

Callado, A., Escribà, P., García-Moya, J.A., Montero, J., Santos, C., Santos-Muñoz, D., and Simarro, J. (2013). Ensemble Forecasting, Climate Change and Regional/Local Responses, Dr Pallav Ray (Ed.), ISBN: 978-953-51-1132-0, InTech, DOI: 10.5772/55699. Disponible en: <http://www.intechopen.com/books/climate-change-and-regional-local-responses/ensemble-forecasting>

Ferranti, L. and Corti, S. (2011): New clustering products, ECMWF Newsletter 127, 6-11. Disponible en : <http://old.ecmwf.int/publications/newsletters/pdf/127.pdf>

Persson, A. and Andersson, E. (2013). User guide to ECMWF forecast products. ECMWF, Reading. Disponible en: http://old.ecmwf.int/products/forecasts/guide/user_guide.pdf