How accurate are operational dust models in predicting Particulate Matter (PM) levels in the Eastern Mediterranean Region? Insights from PM Surface Concentrations

Andreas Eleftheriou[1], Petros Mouzourides[1], Panayiotis Kouis[2], Nikos Kalivitis[3], Itzhak Katra[4], Emily Vasiliadou[5], Chrysanthos Savvides[5], Panayiotis Yiallouros[2], Marina K.-A. Neophytou[1]

[1]Environmental Fluid Mechanics Laboratory, Department of Civil and Environmental Engineering, University of Cyprus, Nicosia, Cyprus

[2]Medical School, Cyprus University, Cyprus

[3]Department of Chemistry, University of Crete, Greece

[4]Department of Environmental, Geoinformatics, and Urban Planning Sciences, Ben-Gurion University of the Negev, Israel

[5]Department of Labour Inspection, Ministry of Labour and Social Insurance, Cyprus

Correspondence to: Petros Mouzourides (pmouzou@ucy.ac.cy)

Abstract. This study provides the first comprehensive assessment of eleven operational dust forecast models and a Multi-model Median (MMM) ~~multi-model ensemble median~~ in predicting ground-level Particulate Matter (PM) concentrations in the Eastern Mediterranean Region (EMR), with a focus on Cyprus, Greece, and Israel. Ground-based observations from regional background stations support model performance assessment across different PM fractions (PM_{10}, $PM_{2.5}$, and coarse particles), using established statistical metrics (correlation coefficient, R, Mean Bias, MB, and Root Mean Square Error, RMSE). ~~The results reveal substantial variability in accuracy, with R values ranging from −0.24 to 0.91 depending on site and event subset. NASA-GEOS consistently achieves the highest correlation (R = 0.71 at Cyprus), indicating accurate representation of dust transport. In contrast, SILAM and EMA-REG4 perform poorly, with low correlations (R = 0.10 and −0.24, respectively) and significant estimation errors (MB = −90.34 μg/m³ for EMA-REG4). The NOA-WRF model effectively captures extreme dust events, with R = 0.91 during the 95th percentile of PM concentrations in Greece. Most models perform better for coarse PM, with the BOOT methodology indicating reduced scatter and bias during dust storm days~~ Across the evaluation subsets, full period, high-dust conditions (95th percentile), and confirmed dust-day events, model skill varies considerably by site and particles fraction (R ranging from −0.24 to 0.91). NASA-GEOS exhibits the highest overall correlations across sites (R = 0.71, 0.65, and 0.64 at Ayia Marina, Be'er Sheva, and Finokalia, respectively), while NOA-WRF demonstrates ~~superior~~ better performance during intense dust events, achieving R = 0.91 at Finokalia (95th percentile). The MMM reduces biases and errors relative to many individual models, providing stable performance across subsets. Performance generally improves for the coarse particle fraction, and BOOT and categorical diagnostics indicate reduced scatter and clearer hit rates during dust-dominated days. However, no single model performs optimally across all sites and conditions;~~,~~ spatial heterogeneity (proximity to sources/orography) and configuration choices (vertical resolution, first-layer height, horizontal resolution, and data assimilation) drive these differences~~highlighting the need for location-specific tuning and evaluation~~. The study underscores the importance of refining model configurations and improving parameterizations to enhance forecast accuracy. Future efforts should incorporate localized data and further develop region-specific models to improve the operational use of these systems in early warning protocols for mitigating public health impacts.

1. Introduction

Desert Dust Storm (DDS) events have been documented since ancient times, with early references found in texts such as "Histories" by Herodotus (Herodotus III: 86-88). These events impact human health and infrastructure, yet they remain understudied until recent decades.

Over the past few years, research into DDSs has expanded rapidly. Recognising their global impacts, the United Nations General Assembly designates 2025-2034 as the Decade for combatting DDSs (United Nations, 2024). This growing attention reflects the mounting evidence linking DDSs to respiratory and cardiovascular illnesses (Lorentzou et al., 2019) and mental health effects (Jones, 2023), regardless of the source region (Lwin et al., 2023). In parallel, climate change, population growth, and geopolitical instability increasingly exacerbate desertification, driving more frequent and intense DDSs and reinforcing transboundary dust transport patterns (Eleftheriou et al., 2023). Therefore, monitoring both the source and downwind impacts of DDSs remains critical for assessing and mitigating their effects.

To address air quality concerns, European directives (e.g. 2004/107/EC, 2008/50/EC, and 2015/1480/EU) and the World Health Organization (WHO) air quality guidelines (2021), establish threshold values to regulate pollution from both anthropogenic and natural sources. These regulatory frameworks enable authorities to track elevated PM concentrations from human activities, while also capturing high PM levels during natural events such as DDSs. During such episodes, PM concentrations often exceed 150–200 $\mu g/m^3$, well above regulatory limits. This capability supports timely public warnings and facilitates health-protective measures when air quality deteriorates due to dust intrusions. The integration of monitoring networks with operational forecast models, such as METAL-WRF (Solomos et al., 2023), enhances our understanding of DDS dynamics and supports compliance with air quality regulations. These systems also strengthen community resilience, providing predictive capacity to mitigate both anthropogenic and natural air pollution. In this context, Early Warning Systems (EWSs) play a key role in protecting the public during hazardous natural events, including DDSs.

EWSs play a critical role in protecting populations from climate-related and natural hazards, including DDSs. EWSs help reduce mortality and economic losses associated with extreme weather and hydrometeorological events. However, major implementation gaps persist, particularly in vulnerable regions such as small island developing states and least-developed countries. According to the United Nations, only 50% of countries operate adequate multi-hazard EWSs, despite the fact that 70% of disaster-related deaths over the past 50 years occurred in the most vulnerable areas. Bridging this gap by advancing forecasting technologies and ensuring broader access to EWSs remains essential, especially for managing transboundary phenomena like DDSs, where timely and accurate forecasts can significantly reduce health and environmental impacts.

The accuracy of EWSs ultimately depends on the performance of dust forecasting models, which vary considerably in their configuration and resolution. These models differ in first-layer calculation heights, ranging from approximately 20 to 100 meters above ground or sea level- and in the particle size ranges they consider, spanning 0.03 to 20µm (Knippertz and Stuut, 2014). They also incorporate diverse emission schemes (e.g., Marticorena and Bergametti, 1995) and deposition processes (e.g., Zender et al., 2003). These differences directly influence model outputs and contribute to substantial variability in dust concentration forecasts.

Model evaluations commonly ~~often~~ rely on sun photometers (e.g., AERONET; Holben et al., 1998) or satellite-derived aerosol optical depth (AOD) products (e.g., MODIS; NASA, 2024), which primarily capture ~~atmospheric~~ column-integrated dust loads rather than near-surface concentrations. Over time, satellite detection algorithms have ~~advanced~~ significantly improved, incorporating geostationary sensors ~~satellites~~ in addition to traditional polar-orbiting platforms ~~ones~~ (e.g., Kolios and Hatzianastassiou, 2019). Unlike polar-orbiting satellites, geostationary platforms provide continuous temporal coverage and are not constrained by the same revisit-time limitations, thereby enhancing the monitoring of dust transport and diurnal variability~~offer improved temporal resolution and are not subject to the same spatial and temporal limitations~~. These approaches, while valuable, are subject to inherent ~~suffer from key~~ limitations. For example, cloud cover can significantly reduce the accuracy of satellite

retrievals, while the low temporal resolution of polar-orbiting satellites constrains their ability to capture rapid dust transport dynamics and event-scale variability ~~reduces satellite retrieval accuracy, and the low temporal resolution of polar-orbiting satellites limits event-scale analysis~~ (Kazadzis et al., 2009). Nevertheless, ongoing initiatives, ~~Despite these challenges, projects~~ such as the Horizon Europe CiROCCO project, are designed to address ~~initiative aim to overcome~~ these limitations ~~issues~~ by developing integrated~~ing~~ predictive frameworks and improving both the spatial and temporal coverage of dust monitoring and forecasting systems~~enhancing dust monitoring and forecasting capabilities~~.

Evaluating dust surface concentrations forecasts using ground PM_{10} concentrations is also challenging, as it is difficult to isolate the desert dust fraction in these measurements (Garcia-Castrillo & Tarradellas, 2017). Consequently, very few studies have evaluated numerical models using data from near-ground monitoring stations, and, to the best of our knowledge, none have focused on the Eastern Mediterranean Region (EMR). This region lies at the crossroads between Africa and Middle East, both of which are major sources of transboundary dust pollution. The study of the EMR has become increasingly difficult in the past decade due to ongoing conflicts and socio-political instability in the nearby dust source regions. This instability has limited in-situ data collection and delayed any mitigation efforts for DDSs (Eleftheriou et al., 2023).

In this study, we evaluate surface dust concentration forecasts from eleven (11) operational dust models and their multi-model median ~~multi-model ensemble median~~ (~~Multi-Model Median;~~MMM ~~MMM~~~~MEDIAN~~) using daily ground-level PM measurements from three background stations in the EMR. These stations, Ayia Marina (AM; Cyprus), Finokalia (FKL; Greece), and Be'er Sheva (BS; Israel), offer low-background environments, allowing for reliable assessment of long-range transboundary dust transport. Observed PM data are categorized into PM_{10}, $PM_{2.5}$, and coarse particle fractions, and we apply a suite of statistical metrics to assess model performance. The measurements include 24-hour averages of both observed PM and modelled dust surface concentrations, ensuring temporal alignment. The performance evaluation incorporates different subsets of the data, including the entire study period, the 95th percentile of PM concentrations, and dust storm days, as classified using the methodology of Achilleos et al. (2020). Section 2 describes the EMR context and the selected study sites. Section 3 outlines data sources, model configurations, and the statistical evaluation methods. Section 4 presents results across all evaluation scenarios and PM fractions, including graphical diagnostics such as Taylor diagrams, BOOT methodology, and contingency tables. Section 5 summarizes findings and discusses implications for model refinement and operational forecasting.

The aim of this work is to determine how accurately current operational models forecast surface dust levels in the EMR, using background PM observations as reference. To our knowledge, this represents the first large-scale, multi-model evaluation of its kind for the region, which is critically exposed to dust intrusions but remains underrepresented in model validation studies. The findings highlight key model strengths and limitations and inform future efforts to improve dust prediction in support of health-focused EWS.

2. Characterizing DDSs in the EMR: Sources and Monitoring

The Mediterranean Basin frequently experiences DDSs, particularly in its southern and eastern regions, which are strongly influenced by emissions from the Sahara and the Sahel (Querol et al., 2009). Recent studies employing combined satellite and ground-based observations have provided a more detailed picture of these events. For instance, Gkikas et al. (2016), using an algorithm that integrates multi-sensor satellite data with in-situ measurements, found that strong dust storms occur more frequently throughout the year in the western Mediterranean Basin, whereas the most intense events tend to develop in the central region. The study also identified a clear seasonal pattern: dust storms are more common during summer in the western Mediterranean, while they predominantly occur in spring across the central and eastern parts. Earlier work by Gkikas et al. (2013) similarly demonstrated that the spatiotemporal characteristics of Mediterranean dust storms are more effectively captured through satellite-based retrievals,

emphasizing the key role of remote sensing in characterizing regional dust dynamics. In addition, Salvador et al. (2022) reported a statistically significant upward trend in the occurrence of African Dust Outbreaks and their intensity over the period 1948–2020. These findings indicate an increase in the frequency of air mass transport from North Africa toward the western Mediterranean Basin and a corresponding intensification of dust episodes during recent decades.

Using air parcel back-trajectory ~~models~~analysis, Varga et al., (2014) further showed ~~demonstrate~~ that dust frequently enters ~~often enters~~ the Eastern Mediterranean Region (EMR) from North~~ern~~ Africa, highlighting the dominant role of African sources in regional dust intrusions. Numerous studies report significant air quality degradation during these events in areas such as mainland Greece, the North Aegean, Cyprus, Israel, and Turkey (Çapraz et al., 2021; Triantafyllou et al., 2020; Tsiflikiotou et al., 2019; Vratolis et al., 2019; Krasnov et al., 2016; Mouzourides et al., 2015). ~~. The more intense events with an extremely high AOD are also documented by Gkikas et al. (2016) in the east Mediterranean whilst the western part receives more frequent dust storms.~~

In addition to African sources, dust emissions from the Middle East significantly affect the EMR. For example, Bodenheimer et al. (2019) analysed 53 DDS~~s~~ events, with concentrations exceeding 150 µg/m³ using data from 13 air quality monitoring stations in Israel (2007–2013). Their findings confirm the transboundary and multi-source nature of dust episodes, which often combine contributions from North African and Middle Eastern source regions.

Figure 1~~Figure 1~~ illustrates the study area and background monitoring sites. These include Ayia Marina Xyliatou (AM) in Cyprus and Finokalia (FKL) in Crete, both part of the European Monitoring and Evaluation Programme (EMEP), and Be'er Sheva (BS) in Israel, part of the Israeli National Air Monitoring Network. These stations are strategically located in relatively isolated areas, minimizing local anthropogenic influence and making them ideal for evaluating long-range dust transport across the EMR.

Each station continuously records PM concentrations using high temporal resolution instruments: the Tapered Element Oscillating Microbalance (TEOM) for Cyprus and Israel, and the FH 62 I-R Thermo analyser for Greece. Measurements are collected at heights of 2–5 m above ground level at AM and FKL, and 10–15 m above ground at BS, where instruments are installed on building rooftops. These setups ensure consistent and representative sampling of ambient air, particularly for coarse PM fractions, which are most affected by dust intrusions. Furthermore, monitoring stations record particle sizes of PM_{10} and $PM_{2.5}$.

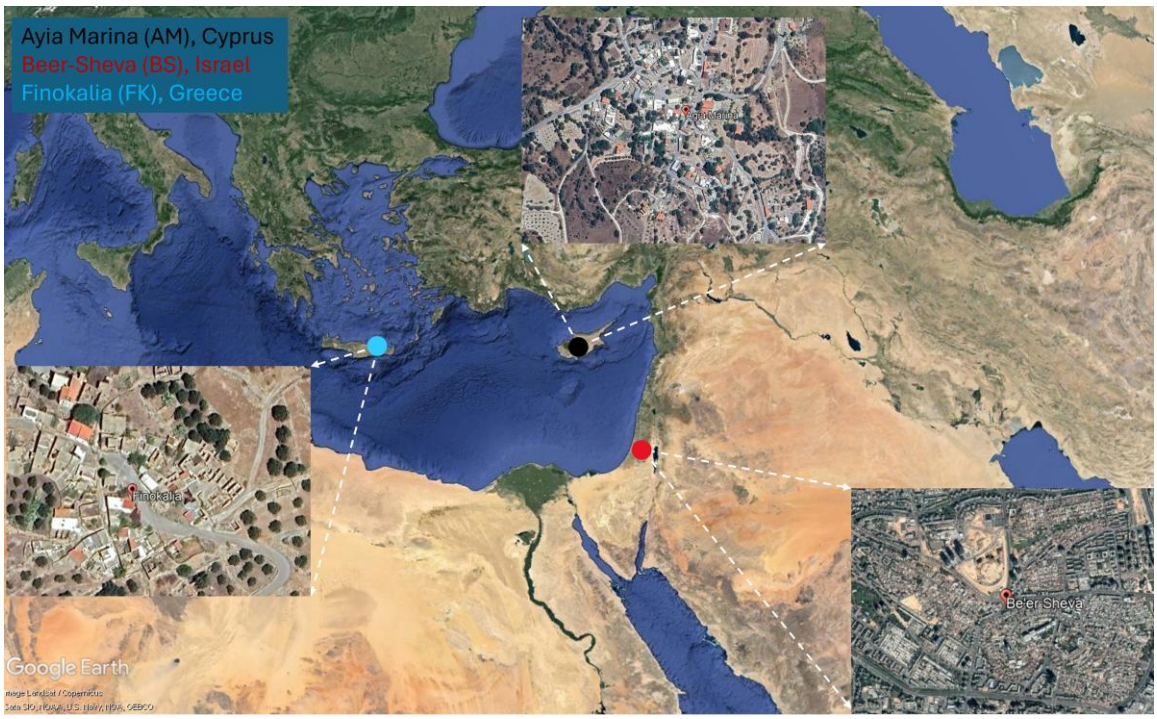


Figure 1~~1~~: ~~Map of the~~S ~~s~~tudy area and monitoring sites in the Eastern Mediterranean. ~~showing the~~L ~~l~~ocations of ~~the three background air quality monitoring stations:~~ Ayia Marina Xyliatou (AM; ~~) in~~ Cyprus) , Finokalia (FKL; ~~) in Crete,~~ Greece), and

Be'er Sheva (BS;) in Israel) used in the model-observation evaluation. The stations are strategically positioned away from major anthropogenic emission sources, making them suitable for evaluating transboundary desert dust transport. Imagery is obtained from © Google Earth 2025.

In this study, we analyse daily averaged forecasts from 11 operational dust models available through the SDS-WAS platform, now known as the WMO Barcelona Dust Regional Center (https://sds-was.aemet.es/). Additionally, we use a MMM Multi-model Median ensemble median (MMMMEDIAN)ensemble, which represents the median surface dust concentration predicted by all participating models for each site. Table 1Table 1 and Table 2 summarise the main configuration characteristics of each model, each model's key configurations like including meteorological drivers, emission schemes, and spatial resolutionconfiguration, providing a consistent basis for performance intercomparisons.

We compare modelled surface dust concentrations to daily observations of PM_{10} and $PM_{2.5}$ collected at the three sites. Before comparing with model outputs, the ground-based PM measurements underwent additional processing to ensure data consistency and reliability. This involves applying quality control checks to remove outliers and erroneous values, converting high-frequency data into 24-hour means, and harmonizing timestamps across sites to correct for time zone and logging differences. These steps are essential for minimizing bias and enabling a robust model–observation comparison.

The observation period covers 2012–2018 for AM, 2012–2017 for BS, and 2012–2016 for FKL. $PM_{2.5}$ data are not available for FKL during this timeframe. Data collection and processing follow the MEDEA project protocol (Achilleos et al., 2020; 2023), co-funded by the LIFE 2016 Programme, which aims to evaluate indoor exposure reduction interventions during DDSs. As part of this effort, Achilleos et al. (2020) develop a dust classification methodology that combines satellite imagery, $PM_{10}/PM_{2.5}$ ratios, elemental composition (Ca, Al, Fe), and dust aerosol optical depth (Dust-AOD) to standardize dust storm identification across all three sites. Using this multi-criteria approach, the authors identify 106 dust storm days at AM, 88 at FKL, and 101 at BS.

Figure 2Figure 2 illustrates presents the time series of daily PM_{10} evolution at concentrations forthe all three regional-background stations, with colour coding (black for AM with black, blue for FKL with blue, red for BS with red) which serves as the basis for the subsequent performance analysesto distinguish each location. These records form the basis for assessing model performance across baseline, extreme, and dust-identified events, as detailed in subsequent sections.The overall concentration levels are relatively low throughout the study period, generally below 50 µg/m³, confirming the regional background character of these sites. Intermittent high-concentration episodes do occur at all locations. Be'er Sheva exhibits the highest variability, with several sharp peaks exceeding 600 µg/m³. At AM, moderate peaks reach up to ~400–500 µg/m³. Finokalia, in contrast, displayed the lowest mean concentrations and the most stable temporal trend, with minimal outlier behaviour. Overall, the three stations represent consistent regional background conditions for the Eastern Mediterranean Basin yet exhibit clear spatial contrasts in the intensity and frequency of episodic dust intrusions.

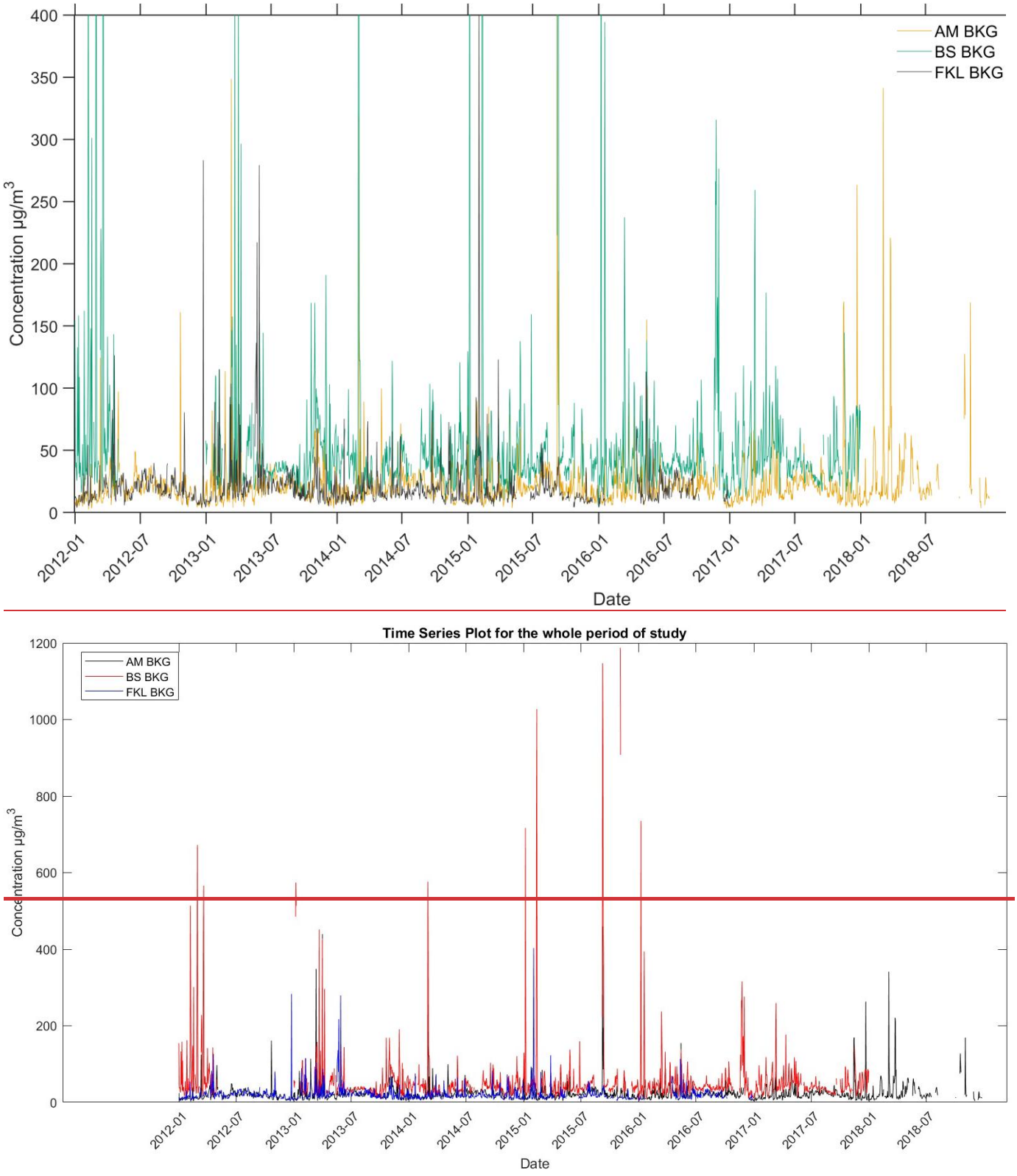


Figure 2: ~~Time series of~~D ~~d~~aily mean PM_{10} concentrations (µg/m³) at the three monitoring stations used for model evaluation, 2012–2018: Ayia Marina, Cyprus (AM, orange), Be'er Sheva, Israel (BS, teal), and Finokalia, Crete (FKL, grey).~~over their respective study periods. Black represents Ayia Marina (Cyprus), blue represents Finokalia (Greece), and red represents Be'er Sheva (Israel).~~ The series show predominantly low background levels (<50 µg m^{-3}) punctuated by episodic high-dust events; more frequent and intense at BS and intermittent at AM, while FKL remains comparatively stable.~~These measurements are used to evaluate dust model performance across baseline and extreme dust conditions.~~

Table 1: Configuration details of the 11 dust forecast models included in this study. These configurations correspond to the versions used during the study period (2012–2018) and may differ from current operational settings.

Model	Institution	Meteorological driver	Regional or global coverage	Meteorological initial fields	Radiation interactions	Emission scheme
BSC_DREAM 8b_V2	BSC	Eta/NCEP	Regional	NCEP/GFS	Dust–radiation feedbacks~~Yes~~	Uplifting (Shao et al., 1993; Janjic et al., 1994; Nickovic ~~Ginoux~~ et al., 2001)
MACC-ECMWF	CAMS	ECMWF	Global	ECMWF/IFS	No	Uplifting (Ginoux et al., 2001; Morcrette et al., 2008, 2009)
DREAM8-MACC	SEEVCCC	NMME	Regional	ECMWF/IFS	No	Uplifting (Shao et al., 1993; Janjic et al., 1994)
NMMB-BSC	BSC	NMMB/NCEP	Regional	NCEP/GFS	Interactive dust-radiation coupling~~Yes~~	Saltation and sandblasting (Janjic et al., 1994; Marticorena and Bergametti, 1995)
NASA_GEOS	NASA	GEOS-5	Global	GEOS-5 Analysis	Direct effects fully included	Based on Ginoux (2001)
NCEP_NGAC	NCEP	NEMS GFS	Global	NCEP/GDAS	Yes (not active)	Dust uplifting following Ginoux (2001)
EMA_REG4	EMA	RegCM4	Regional	NCEP/GFS	Yes (both short and long waves)	Saltation and sandblasting (Zakey et al., 2006; Marticorena and Bergametti, 1995; Alfaro and Gomes, 2001)
DREAMABOL	CNR-ISAC	BOLAM	Regional	NCEP/GFS	Not active	Uplifting (Tegen et al., 1994)
NOA_WRF	NOA	WRF	Regional	GFS	No	GOCART scheme by Ginoux et al. (2001)
SILAM	FMI	ECMWF	Global	Offline Model	Photolysis rates are dependent on overlaying PM in our global 0.2 x 0.2 forecast with chemist	Sofiev et al., 2015
LOTOS	TNO	ECMWF	Regional	ECMWF	No	Marticorena & Bergametti 1995. Dust uplifting following Shao et al. (2001).

Table 2: Output configuration of the dust models used in the performance evaluation. These characteristics influence dust dispersion representation and forecast accuracy at ground level.

Model	Horizontal resolution	Vertical resolution	Height first layer	Transport size bins	Data assimilation	Reference
BSC_DREAM 8b_V2	1/3º x 1/3º	24 Eta-layers	86 m (above sea level)	8 bins (0.1-10µm)	No	Nickovic et al., 2001; Pérez et al., 2006; Basart et al., 2012
MACC-ECMWF	8-10 km approx. (O1280)	137 sigma-layers	10 m (above surface)	3 bins (0.03-20µm)	Yes AOD550/MODIS	Rémy et al., 2022
DREAM8-MACC	1/3º x 1/3º	24 Eta-layers	96 m (above sea level)	8 bins (0.1-10µm)	Yes (ECMWF dust-analysis)	Nickovic et al., 2016
NMMB-BSC	1/3º x 1/3º	24 sigma-hybrid-layers	108 m (above surface)	8 bins (0.1-10µm)	No	Pérez et al., 2011; Klose et al., 2021
NASA_GEOS	0.25º x 0.3125º	72 layers (Top: 0.01 hPa)	Dynamic from 20 to 100 meters-	5 bins (0.73-8.0 µm)	Yes (MODIS)	Colarco et al., 2010
NCEP_NGAC	T126 (~ 1º)	64 sigma-pressure hybrid layers (Top at 0.2 hPa)	20 m	5 bins (0.73-8.0 µm)	No	Lu et al., 2016 Zhang et al., 2022
EMA_REG4	45 km x 45 km	18 sigma-layers	50 m	4 bins (0.01-20)	No	Zakey et al., 2006
DREAMABOL	0.4º x 0.4º	50 sigma-hybrid levels	27 m above surface	8 bins (0.1-10µm)	No	Micrea et al., 2008; Mauruizi et al., 2011
NOA_WRF	0.19º x 0.22º	40 vertical levels	-20	5 bins (0.73-8µm)	No	Flaounas et al., 2017
SILAM	0.1 x 0.1 deg (dust only) 0.2 x 0.2 deg (version with all modeled aerosols and full model chemistry)	19 hybrid levels up to about 16 km (0.1 x 0.1 dust only simulation) 28 hybrid levels up to about 50 km (0.2 x 0.2 simulation)	10m	4 bins (0.1-30 µm)	No	Sofiev et al., 2015
LOTOS	0.50◦longitude × 0.25◦ latitude	12-15 layers	25 m above surface level	5 bins (0.1-10µm)	No	Manders et al., 2017

3. Evaluation Methodology and Performance Metrics

To evaluate the performance of the studied models, we use different statistical metrics, including the Pearson correlation coefficient (R), Mean Bias (MB), and Root Mean Square Error (RMSE). These metrics capture different aspects of model accuracy: R quantifies the strength of the linear relationship between observed and predicted values (ranging from −1 to 1), MB indicates systematic over- or underestimation, and RMSE reflects the overall magnitude of error, with greater sensitivity to large discrepancies. The evaluation is carried out by comparing model outputs with observational data from the three background stations

located in Cyprus, Greece, and Israel. This comprehensive assessment of dust models in the Eastern Mediterranean Region (EMR), to the best of our knowledge has never been done before and provides valuable insights into their performance in an understudied area.

3.1. Statistical Performance Metrics for Model Evaluation

The R coeffcient measures the linear relationship between modelled and observed values, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and a value of 0 implies no correlation ~~(Carslaw, 2015)~~. This metric is crucial for identifying whether the patterns of observation data and forecasts align.

$$R = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{M_i - \bar{M}}{\sigma_M}\right)\left(\frac{O_i - \bar{O}}{\sigma_O}\right) \quad (1)$$

where M_i and O_i are the predicted and observed values respectively, and n is the number of observations.

Mean Bias (MB) is used to determine the average deviation of predicted values from observed values. A positive MB indicates that the model overestimates the observed values, while a negative MB indicates underestimation. This metric is significant as accurate forecasting can help vulnerable populations take precautionary measures during DDS events. The MB is defined as:

$$MB = \frac{1}{n} \sum_{i=1}^{n} M_i - O_i \quad (2)$$

The Root Mean Square Error (RMSE) measures the differences between predicted and observed values and is useful for indicating the model's overall prediction error. RMSE penalizes larger errors more heavily than smaller ones, making it a valuable metric for assessing model performance. It is defined as:

$$RMSE = \left(\sum_{i=1}^{n} \frac{(M_i - O_i)^2}{n}\right)^{1/2} \quad (3)$$

While these metrics effectively summarize model performance, they also have limitations. Specifically, the R and $RMSE$ are linear measures and may fail to capture nonlinear relationships between observed and predicted values. Moreover, although $RMSE$ penalizes larger errors, it remains sensitive to extreme outliers, which potentially skewing the evaluation results, especially during intense dust events. Recognizing these limitations is essential ~~crucial~~ for interpreting the results and identifying areas where models may need further refinement. To enhance the comparative analysis among models, a ranking system was introduced in the performance tables to indicate which models perform best across different sites and statistical metrics.

3.2. Graphical Representations: Taylor Diagrams, BOOT Evaluation and Contingency Tables

Additionally, we employed the Taylor diagram, which provides a graphical summary of the three (3) metrics of R, $RMSE$ and the Standard Deviation (SD) in a single plot. Beyond summarizing performance metrics, the Taylor diagram offers a clear and intuitive means of comparing different models' performances against observational data. In this visualization the SD is proportional to the radial distance from the origin, while $RMSE$ corresponds to the distance from the "observed" point on the x-axis. The green dashed contours indicate specific values and the R is represented along the arc of the diagram, increasing from the top toward the bottom, as illustrated in Figure 3, for example.

The BOOT Model Evaluation Methodology (Chang & Hanna, 2004) is also used, incorporating bootstrapping techniques to resample the data and assess model performance in terms of underestimation or overestimation compared to observed values. The BOOT methodology was chosen over other resampling techniques, such as cross-validation, due to its flexibility and robustness in handling smaller datasets and its ability to provide a statistically sound assessment of model scatter and bias. Bootstrapping allows for the recreation of relationships between observed and modelled data using smaller, resampled datasets (Lahiri, 2010), providing a statistically robust comparison of model performance, and has been used in several recent studies (Lumet et al., 2024;

Yang et al., 2024). This methodology uses two key metrics: Geometric Mean (MG) a measure of relative bias and the Geometric Variance (VG) a measure of relative scatter, with the perfect model positioned at the bottom centre of the graph.

$$MG = exp\left(\overline{lnO_i} - \overline{lnM_i}\right) \quad (4)$$

$$VG = exp\left(\overline{lnO_i - lnM_i}^2\right) \quad (5)$$

We evaluated model skill using categorical statistics that focus on how well models capture observations rather than just their magnitude. Specifically, we built a contingency table that counts, for each day, how often the model overestimates, underestimates, or accurately predicts PM concentrations—where "accurate" means the model falls within ±1 standard deviation of the ground observations. The daily heatmaps use our traffic-light binning for PM levels (0–50, 50–100, 100–200, and >200 µg/m³), and the cell values show the percentage of comparisons in each condition (over, under, within-1SD) relative to the total number of model–observation pairs for that day.~~Finally, categorical statistics were used to evaluate the effectiveness of models in capturing data points, rather than just assessing the intensity of values. A contingency table was developed to display how often models underestimated, overestimated, or accurately predicted PM concentrations (within 1 SD of the ground observations). The numbers in the heatmap represent the percentage of each condition (overestimation, underestimation, or accurate prediction) relative to the total number of model and observation comparisons for each day.~~

3.3. Approaches for Assessing Model Prediction against Observations

In this study, we evaluate model performance using multiple data subsets and comparative approaches to capture different aspects of forecast skill. The observed data include the PM_{10} ($O_{PM_{10}}$), $PM_{2.5}$ (O_{fine}) and PM_{coarse} (O_{coarse}) observations. The latter is calculated as the difference between PM_{10} and $PM_{2.5}$. For the model predictions, only dust concentration data ($M_{dust_{conc}}$) is used. We apply two primary comparison strategies:

i. Modelled dust concentration against PM_{10} observations ($M_{dust_{conc}}$ vs $O_{PM_{10}}$),
ii. Modelled dust concentration against PM_{coarse} observations ($M_{dust_{conc}}$ vs O_{coarse}).

These approaches enable evaluation of total particulate loading and the coarse-mode fraction, which is more directly linked to mineral dust. Each comparison is performed across three temporal evaluation scenarios:

i. the entire study period for each region,
ii. the 95th percentile of observed daily PM concentrations, representing extreme events, and
iii. identified dust storm days, based on the Achilleos et al. (2020), classification framework, which incorporates criteria such as PM_{10} exceedances, Dust Aerosol Optical Depth (Dust-AOD), MODIS satellite imagery, PM ratios, and elemental crustal signatures (Ca, Al, Fe).

This methodology follows the structure proposed by García-Castrillo & Terrandellas (2017) but extends it by incorporating the BOOT evaluation technique (see Section 3.2). To our knowledge, this is the first multi-model validation effort in the Eastern Mediterranean Region (EMR) to apply this combination of methods across multiple PM fractions, locations, and dust-intense periods. For clarity and completeness, we exclude outliers in BOOT visualizations that fall outside the evaluation graph boundaries, as these indicate very low predictive skill and would disproportionately affect interpretability.

4. Assessment of model performance

In this section, we present a comprehensive evaluation of model performance based on the statistical and graphical methods described in Section 3. The results are organized around the three evaluation scenarios: (a) the entire study period, (b) the 95th percentile of observed PM concentrations, and (c) the subset of dust storm days identified using the Achilleos et al. (2020) criteria.

Each model's ability to reproduce observed ground-level PM concentrations is assessed using PM_{10} and O_{coarse} data across all sites. Results are reported using statistical metrics (R, MB, and $RMSE$) and visual diagnostics, including Taylor diagrams, the BOOT methodology, and categorical heatmaps based on contingency analysis. This multi-metric framework allows for a nuanced comparison of model skill, emphasizing both trend representation and bias magnitude under different dust-loading conditions.

4.1. Evaluation of Model Performance Under Different Scenarios

4.1.1. Results for the Entire Study Period

~~**4.1.2. Table 3 presents a summary of model performance across the full study period, based on PM10 and O_{coarse} observations. Below, we present a comparative analysis of individual models, highlighting key trends and site-specific variability.**~~

Table 3 and Figures 3–4 present the comparative evaluation of eleven operational dust forecasting models and the MMM ensemble against observed surface $O_{PM_{10}}$ and O_{coarse} concentrations. Overall, model performance varies substantially among sites, reflecting the combined influence of emission parameterizations, meteorological drivers, and local environmental conditions across the EMR. Most models reproduce the temporal variability of dust concentrations with moderate correlations ($R \approx 0.4 - 0.$~~6~~7) but show systematic biases and site-dependent errors. NASA-GEOS and the MMM ensemble perform best overall, achieving the highest correlations ($R \geq 0.$~~6~~56) and relatively balanced bias profiles. In contrast, EMA-REG4 and ~~LOTUS~~LOTOS-EUROS consistently underperform, with weak correlations ($R \lesssim$~~<~~ 0.3) in most sites and high RMSE values, reaching up to 172 µg/m³ and 144 µg/m³ respectively, at Be'er Sheva.

Distinct spatial patterns emerge across the monitoring network. At AM and FKL, most models underestimate PM_{10} concentrations, with MB values typically between −5 and −15 µg/m³, whereas BS shows consistent overestimation, with MB values frequently between 10 and 40 µg/m³. The highest RMSE values are recorded at BS (70–110 µg/m³), reflecting the site's proximity to active desert dust sources and its strong day-to-day variability. In contrast, lower RMSE values are found at FKL (20–40 µg/m³) and AM (25–35 µg/m³), where observed concentrations represent well-mixed regional background conditions. The best-performing systems, i.e NASA-GEOS, NOA-WRF, and MMM, combine correlation coefficients exceeding 0.6 with relatively low absolute biases ($|MB| \leq 15\ \mu g/m^3$), indicating robust and stable performance across all metrics. Intermediate results are obtained for the DREAM model family (DREAMABOL, DREAM8-MACC, and DREAM8b_V2), with correlations between 0.40 and 0.63, RMSE values between 25 and 85 µg m⁻³, and alternating bias signs depending on site and particle size fraction.

The use of O_{coarse} instead of $O_{PM_{10}}$ generally reduces model bias, particularly at AM, where MB improves from −9.84 µg/m³ to 0.66 µg/m³ for the MMM ensemble and from −13.87 µg/m³ to −3.43 µg/m³ for DREAM8-MACC. However, RMSE values remain within similar ranges (20–80 µg/m³), indicating that coarse-fraction filtering mitigates mean offset but does not eliminate amplitude errors in event representation. BS continues to show the largest spread in MB and RMSE values, with positive MB up to 37 µg/m³ and RMSE exceeding 100 µg/m³, emphasizing the persistent difficulty of simulating near-source dust uplift.

Analysis of the BOOT methodology (Figure 4) supports these conclusions, showing that well-performing models such as NASA-GEOS, NOA-WRF, and MMM cluster closer to the zero-bias line and exhibit reduced scatter, particularly for Be'er Sheva and Ayia Marina. Conversely, low-performing systems like EMA-REG4, ~~LOTUS~~LOTOS-EUROS, and SILAM show points outside the confidence boundaries, indicating substantial deviations in both magnitude and variability. Despite minor inconsistencies between BOOT plots and traditional performance metrics for specific models, the overall trends are consistent, confirming the robustness of the comparative evaluation framework.

In summary, across the full evaluation period, NASA-GEOS and the MMM ensemble are the most consistent performers (R > 0.6, RMSE ≲ 75 μg m⁻³ overall), with NOA-WRF particularly strong at Be'er Sheva (R ≈ 0.7, RMSE = 106 μg m⁻³). EMA-REG4 shows the weakest predictive skill (R ≤ 0.3, RMSE > 170 μg m⁻³), while LOTOS-EUROS exhibits mixed behaviour—performing moderately well at regional background sites but losing skill near dust-source regions. NCEP shows the opposite tendency, correlating better near sources but with positive bias. ~~In summary, across the full period NASA-GEOS and the MMM are the most consistent performers (R > 0.60, RMSE ≲ 75 μg/ m³ overall), with NOA-WRF particularly strong at BS ($R = 0.62$, $RMSE = 106\ \mu g/m^3$). EMA-REG4 and LOTUS show poor predictive skill ($R \leq 0.25$, $RMSE > 140\ \mu g/m^3$).~~ Better agreement at FKL (mean $R \approx 0.55$, $RMSE \approx 25\ \mu g/m^3$) indicates that models reproduce long-range transport more accurately than near-source processes. Overall, the ensemble median (MMM) reduces individual model variability and provides the most balanced representation of $O_{PM_{10}}$ and O_{coarse} concentrations, underscoring its utility for operational dust forecasting in the EMR.

~~DREAM8-MACC achieves moderate correlations, with R = 0.54 at FKL, 0.51 at BS, and 0.48 at AM. It underestimates PM10 at AM (MB = −13.87 μg/m³) and FKL (MB = −13.55 μg/m³) but slightly overestimates at BS (MB = 0.57 μg/m³). RMSE values are 26.05 μg/m³ (AM), 21.93 μg/m³ (FKL), and 79.26 μg/m³ (BS), indicating larger prediction errors in Israel. For O_{coarse}, MB shifts to −3.43 μg/m³ (AM) and 21.84 μg/m³ (BS), suggesting sensitivity to particle fraction. The BOOT methodology is providing a good representation for the BS area with regards to the performance metrics whilst for AM it shows more underestimation than what the performance metrics indicate.~~

~~DREAM8b_V2 also shows moderate correlation (R = 0.58 at FKL, 0.49 at BS, 0.40 at AM), underestimating PM10 at AM (MB = −7.93 μg/m³) and FKL (MB = −10.32 μg/m³), and overestimating at BS (MB = 0.57 μg/m³). RMSE ranges from 24.08 μg/m³ (FKL) to 85.20 μg/m³ (BS). For O_{coarse}, MB increases to 2.60 μg/m³ (AM) and 56.95 μg/m³ (BS), highlighting model sensitivity to coarse particle loading. The BOOT methodology graph aligns with the performance metrics, particularly for AM and BS.~~

~~DREAMABOL performs best at FKL (R = 0.63), with lower values at BS (0.49) and AM (0.42). It slightly underestimates PM10 at AM (MB = −4.04 μg/m³) and FKL (MB = −7.25 μg/m³), while overestimating at BS (MB = 3.66 μg/m³). RMSE is highest at BS (69.35 μg/m³). In the PM_{coarse} comparison, MB increases at AM (6.32 μg/m³) and decreases slightly at BS (23.31 μg/m³). The BOOT methodology for DREAMABOL shows consistency with the performance metrics described above except from the AM situation where even though MB changes from negative to positive between the approaches on the BOOT graph is still on the underestimation side.~~

~~EMA-REG4 shows the weakest performance, with R = 0.31 at FKL, 0.25 at BS, and just 0.02 at AM. It heavily underestimates at AM (MB = −19.54 μg/m³) and FKL (MB = −12.48 μg/m³), but overestimates at BS (MB = 16.35 μg/m³). RMSE peaks at 172.40 μg/m³ (BS). Under O_{coarse}, MB improves at AM (−8.77 μg/m³) but worsens at BS (36.53 μg/m³), while RMSE remains high (177.69 μg/m³). The poor performance of the model is also depicted in the BOOT methodology graph since the points for AM and FKL are out of the boundaries. Also, the points for BS are showing underestimation while the MB is positive.~~

~~LOTOS exhibits low correlation, with R = 0.24–0.47, and consistent underestimation at all sites: MB = −2.74 μg/m³ (AM), −2.63 μg/m³ (BS), −12.33 μg/m³ (FKL). RMSE values are highest at BS (144.44 μg/m³) and AM (46.39 μg/m³). Under O_{coarse}, MB reverses to 8.21 μg/m³ (AM) and 13.91 μg/m³ (BS), indicating a model bias toward larger particle overestimation. The low correlation of the model is also found on the BOOT methodology graph as the points for all sites are out of boundaries except from the BS point of the O_{coarse} approach that shows significant overestimation.~~

~~MACC-ECMWF performs moderately across all sites (R = 0.47–0.54), with negative MBs: −11.84 μg/m³ (AM), −28.50 μg/m³ (BS), −11.51 μg/m³ (FKL). RMSE is highest at BS (72.78 μg/m³). Under O_{coarse}, MB reduces significantly to −1.39 μg/m³ (AM) and −7.83 μg/m³ (BS), suggesting better alignment when fine PM is removed. The BOOT methodology captures the decrease in the negative values between the approaches for both Ayia Marina and Be'er Sheva.~~

~~The Multi-Model-Median (MMM) achieves good correlation overall, with R = 0.64 (FKL), 0.58 (BS), and 0.56 (AM). It tends to underestimate at AM (MB = −9.84 μg/m³) and FKL (MB = −10.88 μg/m³) and slightly overestimate at BS (MB = 13.74 μg/m³). RMSE values are 23.55 μg/m³ (AM), 20.43 μg/m³ (FKL), and 71.79 μg/m³ (BS). Under O_{coarse}, MB at AM improves to near-neutral (0.66 μg/m³), while increasing at BS (34.78 μg/m³). In the BOOT methodology, MMM supports the results in the performance metrics with Ayia Marina O_{coarse} and the Be'er Sheva PM10 approaches to be better performing.~~

NASA GEOS performs best overall, with values of R = 0.71 (AM), R = 0.65 (BS), and R = 0.64 (FKL). These high correlations indicate that NASA GEOS effectively captures the trends in PM10 concentrations at all locations, though the relatively high RMSE values, particularly at Be'er Sheva (89.48 μg/m3), highlight challenges in predicting extreme dust concentrations. The model's strong performance can be attributed to its configuration settings which include detailed representation of global dust dynamics (not regional) and effective calibration for long range dust transport (i.e Data assimilation). In terms of MB, NASA GEOS shows a slight overestimation in Cyprus (MB = 3.60 μg/m³) and Israel (MB = 14.98 μg/m³), while it performs exceptionally well at Crete-Greece, where the bias is nearly zero (MB = 0.64 μg/m³). The model's RMSE values are relatively high, with the largest error seen at Israel (RMSE = 89.48 μg/m³), followed by Greece (RMSE = 41.24 μg/m³) and Cyprus (RMSE = 35.07 μg/m³), indicating some challenges in predicting extreme PM10 concentrations despite its overall strong trend capture. The models perform similarly in both approaches ($O_{PM_{10}}$ and O_{coarse}) but with decreases in R and increases in MB and RMSE for the O_{coarse} dataset for both areas (AM and BS). This difference indicates that the models underperform when trying to predict concentrations of larger particles. In the BOOT methodology analysis, NASA GEOS performs best in both $M_{dust\ conc}$ vs $O_{PM_{10}}$ and $M_{dust\ conc}$ vs O_{coarse} approaches, further highlighting its strong predictive capabilities.

NCEP NGAC exhibits site-dependent performance, with R = 0.62 at FKL, 0.52 at BS, and just 0.27 at AM. MB is positive across sites (3.40–7.96 μg/m³), while RMSE varies from 29.24 μg/m³ (FKL) to 68.76 μg/m³ (BS). O_{coarse} results show elevated MB: 18.34 μg/m³ (AM), 28.91 μg/m³ (BS). The Boot methodology graph presents an a very good performance on all 3 sites for the PM10 approach in comparison with the rest of the models. In terms of the O_{coarse} approach there is a shift towards the overestimation side which is consistent with the results of the performance metrics for Cyprus and Greece.

NMMB BSC records moderate correlation, with R = 0.56 (BS), 0.43 (AM), and 0.42 (FKL). MB is consistently negative: −15.03 μg/m³ (AM), −25.04 μg/m³ (BS), and −14.91 μg/m³ (FKL). RMSE is highest at BS (86.19 μg/m³). In the O_{coarse} scenario, MB improves: −4.46 μg/m³ (AM), −3.68 μg/m³ (BS).

NOA-WRF performs well, particularly at BS (R = 0.62), FKL (0.58), and AM (0.56). MB shows overestimation: 7.41 μg/m³ (AM), 36.54 μg/m³ (BS), 1.67 μg/m³ (FKL). RMSE values reach 105.99 μg/m³ at BS. In O_{coarse} comparisons, MB increases to 17.54 μg/m³ (AM) and 54.64 μg/m³ (BS), consistent with the model's known bias in high dust load conditions. The BOOT methodology analysis shows that NOA-WRF consistently performs well for Be'er Sheva, particularly in the $M_{dust\ conc}$ vs $O_{PM_{10}}$ approach, further validating its robustness despite the tendency to overestimate. On the other hand, the BOOT methodology shows for both Ayia Marina and FKL sites to be always underestimating the observed values.

SILAM performs poorly overall, with R = 0.49 (FKL), 0.40 (BS), and 0.38 (AM). MB is negative at all sites: −7.31 μg/m³ (AM), −8.33 μg/m³ (BS), −9.22 μg/m³ (FKL). RMSE is highest at BS (95.29 μg/m³). Under O_{coarse}, MB becomes positive: 3.14 μg/m³ (AM), 8.73 μg/m³ (BS), suggesting the model underrepresents fine particles. The BOOT methodology is not supporting any of the findings for this model as no points are shown within the boundaries of the graph.

Summarizing the results from the site specific analysis, Ayia Marina exhibits the highest overall correlation values across most models, with NASA GEOS performing particularly well (R = 0.71). NOA-WRF and the MMM also effectively capture dust trends at this site, suggesting that key dust sources are well represented. In contrast, EMA-REG4 and NCEP show substantial limitations, with correlations as low as R = 0.02 and R = 0.27, respectively, indicating poor simulation of dust transport for this region. At Be'er Sheva, model performance varies significantly, with systematic overestimation observed in several models (e.g., NOA-WRF MB = 36.54 μg/m³, NASA GEOS MB = 14.98 μg/m³). This variability likely reflects the site's proximity to major Middle Eastern dust sources, which increases forecast complexity and uncertainty. At Finokalia, models generally show stronger agreement with observations, with NASA GEOS and MMM achieving R = 0.64. The site's relative isolation from anthropogenic pollution makes it well-suited for evaluating long-range transboundary dust transport. However, models like LOTUS continue to underperform (R = 0.23), indicating persistent difficulties in simulating dust concentrations accurately at this location. Comparing the $O_{PM_{10}}$ and O_{coarse} approaches, results do not change substantially in terms of correlation. However, mean bias values fluctuate, with some models shifting from negative to positive MB and vice versa, depending on particle size fraction. These variations highlight uncertainties linked to particle composition and model sensitivity to fine vs. coarse PM. Overall, the findings underscore the need for improved model tuning, especially with respect to local conditions and dominant particle size distributions in the Eastern Mediterranean Region.

For the entire dataset (Table 3; Figures 3–4; Figure 9), both OPM10 and Ocoarse show that NASA GEOS and the multi-model median (MEDIAN) generally matched the observations most closely, though with important differences.

For OPM10, at AM, NASA GEOS gave the highest correlation (R = 0.71) with a small positive bias of +3.60 μg/m³, although the RMSE was relatively high at 35.07 μg/m³. MEDIAN gave a slightly lower correlation of 0.56 but the lowest RMSE at 23.55 μg/m³ and a moderate negative bias of −9.84 μg/m³. At BS, NASA GEOS again gave the highest correlation (R = 0.65) but had a large RMSE of 89.48 μg/m³ and a positive bias of +14.98 μg/m³. MEDIAN produced a correlation of 0.58 with lower RMSE (71.79

~~μg/m³) and smaller bias (+13.74 μg/m³). At FKL, NASA_GEOS and MEDIAN both reached a correlation of 0.64. NASA_GEOS had a higher RMSE (41.24 μg/m³), while MEDIAN had the lowest RMSE at 20.43 μg/m³ and a negative bias of −10.88 μg/m³.~~

~~For Ocoarse, NASA_GEOS produced the highest correlation at AM (R = 0.70) with bias of +14.23 μg/m³ and at BS (R = 0.60) with bias of +36.11 μg/m³. MEDIAN gave more even behaviour with correlations near 0.56, the lowest RMSE at 18.46 μg/m³ for AM, and smaller biases, though still positive at BS (+34.78 μg/m³). DREAMABOL, NMMB, and MACC_ECMWF gave intermediate values. SILAM and EMA_REG4 produced the lowest correlations, in some cases below 0.25, with very large RMSE.~~

~~Figure 9 indicates that for both fractions MEDIAN had the highest share of values within ±1 SD. NASA-GEOS and NOA-WRF tended to overestimate, especially at BS, while EMA_REG4 and SILAM showed the largest scatter.~~

~~In summary, over the entire study period NASA_GEOS gave the highest correlations with OPM10and Ocoarse but often with higher RMSE and positive bias. MEDIAN achieved lower errors and more moderate biases. DREAMABOL compared well with OPM10 at FKL, NOA_WRF was reasonable at BS but overestimated, and EMA_REG4, SILAM, and LOTUS consistently gave the weakest results.~~

Table ~~3~~3: Ranked statistical performance (correlation coefficient *R*, mean bias *MB*, and root mean square error *RMSE*) of the eleven operational dust-forecast models and the MMM against observed PM_{10} ($O_{PM_{10}}$) and coarse fraction (O_{coarse}) at Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL) over the full study period. Table 3a summarizes results for $O_{PM_{10}}$, and Table 3b for O_{coarse}. Rankings are by~~.....~~ performance of R coefficient

~~Statistical performance metrics (correlation coefficient R, mean bias MB, and root mean square error RMSE) for modelled dust concentrations compared to observed PM_{10} and O_{coarse} values at three background stations (Ayia Marina – AM, Be'er Sheva – BS, and Finokalia – FKL) over the full study period.~~Ranked statistics of model performance against PM10 (a) and PMcoarse (b) ~~for the entire study period~~

~~(a)~~

Model	AM_R (R)	AM_RMSE (μg/m³)	AM_MB (μg/m³)		BS_R (R)	BS_RMSE (μg/m³)	BS_MB (μg/m³)		FKL_R (R)	FKL_RMSE (μg/m³)	FKL_MB (μg/m³)
NASA_GEOS~~DREAM8_MACC~~	0.710.48	35.0726.05	3.6~~-13.87~~	NASA_GEOS	0.650.51	89.4879.26	14.980.57	NASA_GEOS	0.640.54	41.2421.93	0.64~~-13.55~~
NOA_WRF~~DREAM8b_V2~~	0.560.40	38.3827.16	7.41~~-7.93~~	NOA_WRF	0.620.49	105.9985.20	36.540.57	MMM	0.640.58	20.4324.08	~~-10.88~~-10.32
MMM~~DREAMABOL~~	0.560.42	23.5527.53	~~-9.84~~-4.04	MMM	0.580.49	71.7969.35	13.743.66	DREAMABOL	0.630.63	19.7219.72	~~-7.25~~-7.25
DREAM8_MACC~~EMA_REG4~~	0.480.02	26.0534.27	~~-13.87~~-19.54	NMMB	0.560.25	86.19172.40	-25.0416.35	NCEP	0.620.31	29.2421.54	3.4~~-12.48~~
LOTOS-EUROS~~LOTUS~~	0.470.47	46.3946.39	~~-2.74~~-2.74	MACC_ECMWF	0.540.24	72.78144.44	~~-28.5~~-2.63	DREAM8b_V2	0.580.23	24.08131.11	~~-10.32~~-12.33
MACC_ECMWF~~MACC_ECMWF~~	0.470.47	26.5626.56	~~-11.84~~-11.84	NCEP	0.520.54	68.7672.78	7.96~~-28.50~~	NOA_WRF	0.580.50	33.0128.08	1.67~~-11.51~~
NMMB~~NASA_GEOS~~	0.430.71	28.1735.07	~~-15.03~~3.60	DREAM8_MACC	0.510.65	79.2689.48	0.57~~14.98~~	DREAM8_MACC	0.540.64	21.9341.24	-13.550.64
DREAMABOL~~NCEP~~	0.420.27	27.5354.03	~~-4.04~~7.88	DREAM8b_V2	0.490.52	85.26~~8.76~~	0.57~~7.96~~	MACC_ECMWF	0.50~~.62~~	28.0829.24	-11.513.40
DREAM8b_V2~~NMMB~~	0.40~~.43~~	27.1628.17	~~-7.93~~-15.03	DREAMABOL	0.490.56	69.3586.19	3.66~~-25.04~~	SILAM	0.490.42	31.9425.54	~~-9.22~~-14.91
SILAM~~NOA_WRF~~	0.380.56	61.4238.38	~~-7.31~~7.41	SILAM	0.40~~.62~~	95.2910~~5.99~~	-8.3336.54	NMMB	0.420.58	25.5433.01	-14.911.67
NCEP~~SILAM~~	0.270.38	54.0361.42	7.88~~-7.31~~	EMA_REG4	0.250.40	172.495.29	16.35~~-8.33~~	EMA_REG4	0.310.49	21.5431.94	~~-12.48~~-9.22
EMA_REG4~~MMM~~	0.020.56	34.2723.55	~~-19.54~~-9.84	LOTOS-EUROS	0.240.58	144.447~~1.79~~	-2.6313.74	LOTOS-EUROS	0.230.64	13.1120.43	~~-12.33~~-10.88

(b)

Model	AM (R)	AM$_{RMSE}$ (µg/m³)	AM$_{MB}$ (µg/m³)		BS (R)	BS$_{RMSE}$ (µg/m³)	BS$_{MB}$ (µg/m³)		FKL (R)	FKL$_{RMSE}$ (µg/m³)	FKL$_{MB}$ (µg/m³)
NASA_GEOSDREAM8_MACC	0.70.47	39.4417.96	14.23-3.43	NOA_WRF	0.670.49	113.9980.00	54.6421.84		–	–	–
NOA_WRFDREAM8b_V2	0.570.41	42.7522.94	17.5422.60	NASA_GEOS	0.60.46	100.0992.51	36.1156.95		–	–	–
MMMDREAMABOL	0.560.42	18.4625.95	0.6666.32	MMM	0.560.48	76.4765.61	34.7823.31		–	–	–
DREAM8_MACCEMA_REG4	0.470.01	17.9624.79	-3.43-8.77	MACC_ECMWF	0.510.24	60.59177.69	-7.8336.53		–	–	–
MACC_ECMWFLOTUS	0.460.45	20.8347.30	-1.398.21	NMMB	0.510.30	85.56143.73	-3.6813.91		–	–	–
LOTOS-EUROSMACC_ECMWF	0.450.46	47.320.83	8.21-1.39	DREAM8_MACC	0.490.51	8060.59	21.84-7.83		–	–	–
DREAMABOLNASA_GEOS	0.420.70	25.9539.44	6.3214.23	NCEP	0.490.60	69.94100.09	28.91136.11		–	–	–
NMMBNCEP	0.420.27	20.7856.16	-4.4618.34	DREAMABOL	0.480.49	65.6169.94	23.3128.91		–	–	–
DREAM8b_V2NMMB	0.410.42	22.9420.78	2.6-4.46	DREAM8b_V2	0.460.51	92.5185.56	56.95-3.68		–	–	–
SILAMNOA_WRF	0.370.57	61.7742.75	3.1417.54	SILAM	0.410.67	94.88113.99	8.7354.64		–	–	–
NCEPSILAM	0.270.37	56.1661.77	18.3434.14	LOTOS-EUROS	0.30.41	143.7394.88	13.918.73		–	–	–
EMA_REG4MMM	0.010.56	24.7918.46	-8.770.66	NOA_WRF	0.670.56	113.9976.47	54.6434.78		–	–	–

(b)

Model	AM_R (R)	AM_RMSE (µg/m³)	AM_MB (µg/m³)	BS_R (R)	BS_RMSE (µg/m³)	BS_MB (µg/m³)	FKL_R (R)	FKL_RMSE (µg/m³)	FKL_MB (µg/m³)
DREAM8_MACC	4 (0.47)	1 (17.96)	5 (-3.43)	6 (0.49)	5 (80.00)	5 (21.84)	-	-	-
DREAM8b_V2	9 (0.41)	5 (22.94)	3 (2.60)	9 (0.46)	7 (92.51)	12 (56.95)	-	-	-
DREAMABOL	7 (0.42)	7 (25.95)	7 (6.32)	8 (0.48)	2 (65.61)	6 (23.31)	-	-	-
EMA_REG4	12 (0.01)	6 (24.79)	9 (-8.77)	12 (0.24)	12 (177.69)	10 (36.53)	-	-	-
LOTUS	6 (0.45)	10 (47.30)	8 (8.21)	11 (0.30)	11 (143.73)	4 (13.91)	-	-	-
MACC_ECMWF	5 (0.46)	4 (20.83)	2 (-1.39)	4 (0.51)	1 (60.59)	2 (-7.83)	-	-	-
NASA_GEOS	1 (0.70)	8 (39.44)	10 (14.23)	2 (0.60)	9 (100.09)	9 (36.11)	-	-	-
NCEP	11 (0.27)	11 (56.16)	12 (18.34)	6 (0.49)	3 (69.94)	7 (28.91)	-	-	-
NMMB	7 (0.42)	3 (20.78)	6 (-4.46)	4 (0.51)	6 (85.56)	1 (-3.68)	-	-	-
NOA_WRF	2 (0.57)	9 (42.75)	11 (17.54)	1 (0.67)	10 (113.99)	11 (54.64)	-	-	-
SILAM	10 (0.37)	12 (61.77)	4 (3.14)	10 (0.41)	8 (94.88)	3 (8.73)	-	-	-
MEDIAN	3 (0.56)	2 (18.46)	1 (0.66)	3 (0.56)	4 (76.47)	8 (34.78)	-	-	-

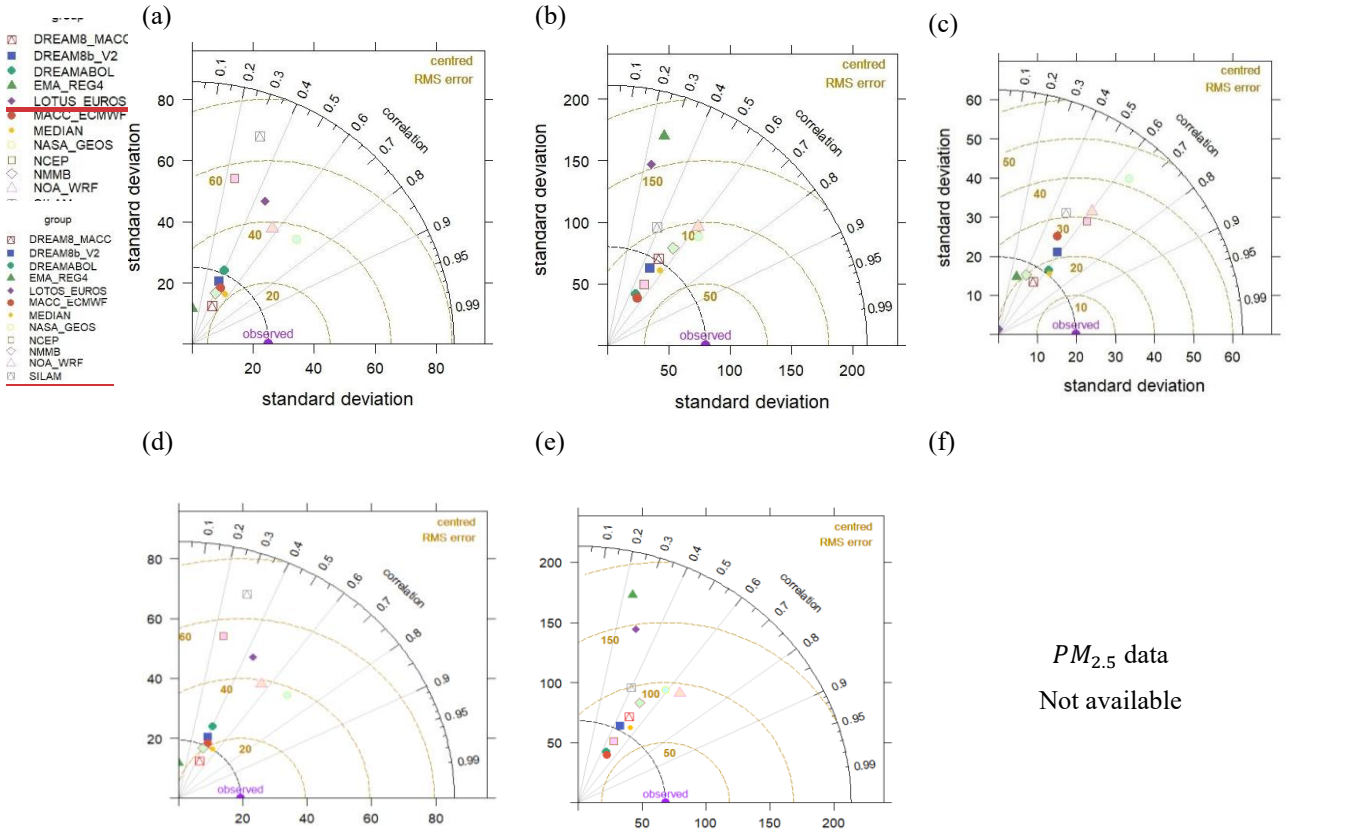
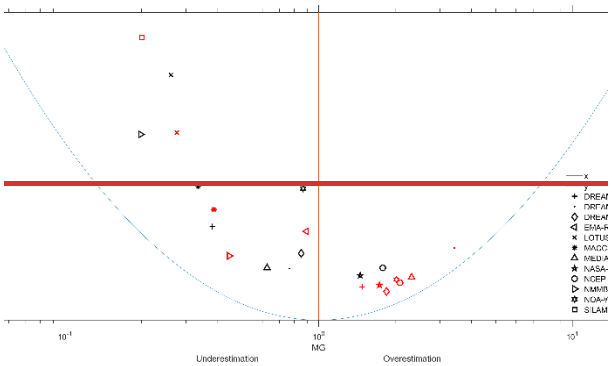


Figure 3: ~~Taylor diagrams showing the performance of 11 operational dust models and the multi-model median (MMMMEDIAN) against observed daily PM_{10} concentrations during the full study period at the three monitoring sites: (a and c) Ayia Marina (AM), (b and e) Be'er Sheva (BS) and (c and f) Finokalia (FKL).~~Taylor diagrams for daily PM_{10} over the full study period at the three monitoring sites. Panels show model performance against observations at (a) Ayia Marina (AM), (b) Be'er Sheva (BS), and (c) Finokalia (FKL). The azimuth indicates the correlation coefficient (R), the radial distance the standard deviation (SD), and the concentric arcs the centered RMSE. Symbols denote the 11 operational models; the multi-model median (MMM) is highlighted. The observational reference is located at R=1 and SD=1.

(a)

(b)



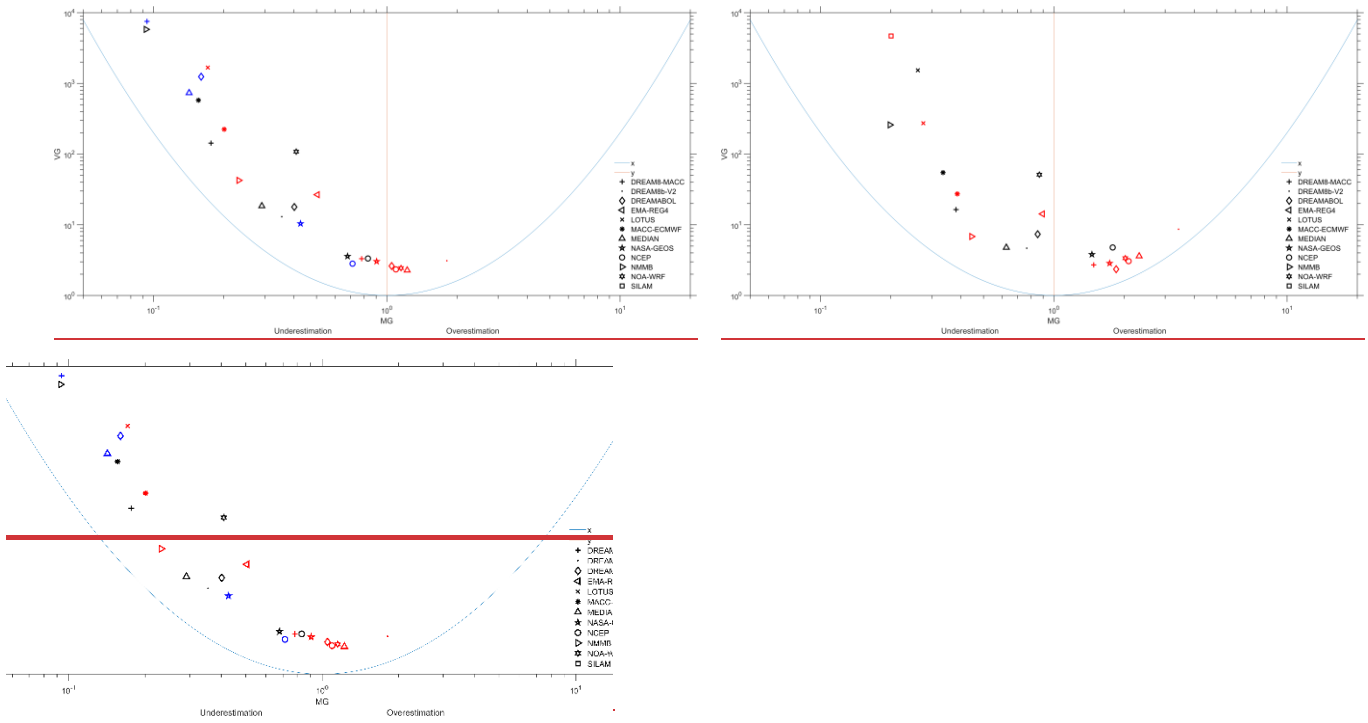


Figure 4: BOOT statistical plots comparing modelled daily PM_{10} concentrations to observed values over the full study period at the three monitoring stations: Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). a) $M_{dust_{conc}}$ vs $O_{PM_{10}}$ and b) $M_{dust_{conc}}$ vs O_{Coarse}. Models closer to the origin demonstrate better agreement with observations. Outliers beyond graph limits are excluded for visual clarity.

~~4.1.3.~~4.1.2. Results for the 95th Percentile of Measurements

~~This evaluation focuses on the models' ability to simulate extreme dust events, represented by the 95th percentile of daily mean PM concentrations. Results are presented for $O_{PM_{10}}$ and O_{coarse}, using correlation (R), mean bias (MB), and RMSE across the three monitoring stations.~~

This analysis evaluates model performance in reproducing the most intense dust events, represented by the 95th percentile of daily mean PM concentrations. Overall, most models exhibit reduced skill compared to the full-period evaluation, highlighting the challenge of simulating extreme events where both emissions and transport processes become highly nonlinear. Nevertheless, clear contrasts among models and sites emerge. NASA-GEOS and NOA-WRF consistently rank among the best-performing systems, achieving the highest correlations and accurately capturing the timing of peak dust episodes across the EMR. NOA-WRF reaches R = 0.91 at FKL, the highest correlation recorded and maintains good agreement at AM (R = 0.53) and BS (R = 0.~~62~~55). Similarly, NASA-GEOS performs strongly at AM (R = 0.66) and FKL (R = 0.56)~~BS (R = 0.65)~~, with slightly lower skill at BS (R = 0.49). ~~FKL (R = 0.49).~~ These results indicate that both models effectively reproduce the temporal evolution of major dust outbreaks, particularly in coastal and inland environments strongly influenced by African and Middle Eastern sources. However, their positive mean biases, especially at BS (up to 150–200 µg/m³), demonstrate a systematic overestimation of event intensity. This behaviour suggests that while these models capture the occurrence of dust storms well, they overpredict near-source emissions and coarse-particle loadings, resulting in elevated RMSE values exceeding 300 µg/m³.

A second group of models, including DREAMABOL, DREAM8-MACC, DREAM8b_V2, NCEP, and the MMM ensemble, show moderate skill, with correlations typically between 0.4 and 0.6 and mixed bias behaviour. For example, DREAMABOL and DREAM8b_V2 reproduce high-dust variability at FKL and BS but underestimate at AM, while NCEP and the MMM show more balanced bias patterns, with moderate underestimation in Cyprus and mild overestimation in Israel. The MMM continues to provide robust, smoothed estimates of dust concentrations, reducing extremes and offering a stable compromise among individual model

tendencies. Despite these advantages, RMSE values remain high (often above 250 μg/m³), revealing that even the ensemble approach struggles to match the magnitude of extreme dust concentrations.

In contrast, models such as EMA-REG4, NMMB, and SILAM show limited ability to reproduce severe dust events, with correlations typically below 0.3–0.4 and RMSE exceeding 250–450 μg m⁻³. LOTOS-EUROS performs similarly overall but shows localized improvements, achieving R = 0.59 and RMSE = 54 μg m⁻³ at Ayia Marina, indicating occasional skill in long-range transport under specific high-dust conditions. ~~In contrast, models such as EMA-REG4, LOTOS, NMMB, and SILAM demonstrate limited ability to capture the dynamics of severe dust events. These systems generally yield low or even negative correlations (R < 0.3) and substantial errors, with RMSE values exceeding 300–450 μg/m³.~~

The comparison between $O_{PM_{10}}$ and O_{Coarse} simulations, provides additional insight. For many models, bias decreases when fine particles are excluded, indicating that misrepresentation of the fine fraction contributes significantly to the total PM_{10} error. For instance, DREAM8-MACC at Ayia Marina improves from MB = −54.97 to −28.81 μg/m³, and NCEP at BS from −86.47 to −22.56 μg/m³ when using O_{Coarse}. However, improvements in RMSE are limited, showing that large errors in event magnitude persist even after accounting for size partitioning. These findings confirm that reproducing the intensity of extreme dust episodes remains a critical challenge, as current emission and transport schemes struggle to accurately represent the rapid onset and decay of strong outbreaks.

~~NOA-WRF and NASA-GEOS rank as the top-performing models based on their consistently higher correlation coefficients and their ability to reproduce the temporal variability of extreme dust concentrations. Specifically, NOA-WRF achieves R = 0.91 at Finokalia, the highest correlation across all models and stations, reflecting its strong capacity to track peak dust events at this site. It also performs well at Ayia Marina (R = 0.53) and Be'er Sheva (R = 0.62), indicating robust spatial consistency. Similarly, NASA-GEOS shows strong correlations at Ayia Marina (R = 0.66) and Be'er Sheva (R = 0.65), and moderate performance at Finokalia (R = 0.49). These values suggest that NASA-GEOS more accurately represents the magnitude and timing of intense dust episodes at inland and island locations. Despite these strengths in correlation, both models exhibit notable positive bias, particularly at Be'er Sheva, where NOA-WRF and NASA-GEOS overestimate PM_{10} by 197.87 μg/m³ and 153.10 μg/m³, respectively, and show high RMSEs (302.59 and 300.83 μg/m³). These results indicate that while the models successfully capture dust event occurrence, they tend to exaggerate intensity, especially in proximity to dust source regions. The results from both models were supported by the BOOT methodology, which confirmed their relatively strong predictive capability, particularly in terms of capturing temporal trends (correlation), despite persistent overestimation during high dust events.~~

~~A group of moderate performers includes DREAMABOL, DREAM8-MACC, DREAM8b_V2, NCEP, and the MMM ensemble. These models exhibit moderate correlation values, typically ranging from R = 0.20 to 0.55. For instance, DREAMABOL achieves R = 0.60 at FKL, while DREAM8b_V2 records R = 0.45 at BS. Biases vary: DREAMABOL and DREAM8b_V2 show underestimations at AM (MB = −24.00 and −42.32 μg/m³, respectively), while NCEP shows mixed performance, with MB = −86.47 μg/m³ at BS, but +21.92 μg/m³ at FKL. The MMM model yields R = 0.55 at FKL and relatively balanced MB values (e.g., −8.59 μg/m³ at AM, 17.00 μg/m³ at BS), but RMSE remains high across all locations (e.g., 252.92 μg/m³ at BS).~~

~~Models with consistently low performance were EMA-REG4, LOTUS, NMMB, and SILAM. EMA-REG4 showed R = −0.24 at AM and large underestimation (MB = −90.34 μg/m³). At BS, despite R = 0.41, RMSE was 447.58 μg/m³, the highest among all models. LOTUS produced weak correlations across all sites, except in O_{coarse} at AM (R = 0.59). MB ranged from −20.79 μg/m³ to +98.75 μg/m³. NMMB systematically underestimated (MB = −50.52 to −67.19 μg/m³) with poor correlation (R < 0.4) and RMSE up to 315.38 μg/m³. SILAM performed inconsistently, with R = 0.10 at AM, and although MB was near zero at FKL (1.48 μg/m³), RMSE exceeded 244 μg/m³. Among sites, Finokalia showed the best overall agreement, especially for NOA-WRF (R = 0.91), DREAMABOL (0.60), and MMM (0.55). Ayia Marina had high R for NASA-GEOS (0.66), but very poor scores for EMA-REG4 (R = −0.24). Be'er Sheva was the most challenging, with most models overestimating PM_{10} and RMSEs frequently exceeding 250 μg/m³. Using O_{coarse} instead of PM_{10} generally improved MB for several models. For instance, DREAM8-MACC at AM improved from MB = −54.97 to −28.81 μg/m³, and NCEP at BS from −86.47 to −22.56 μg/m³. However, RMSE reductions were limited, showing that errors in event intensity estimation persist even when fine particles are excluded.~~

~~In conclusion, while a few models (notably NOA-WRF and NASA-GEOS) demonstrated strong potential for predicting high-dust events, most models struggled with significant biases and poor precision. The coarse fraction approach slightly improved MB in many cases but did not resolve overall performance limitations. These results highlight the need for improved model representation of dust sources and transport, especially in complex environments like BS.~~

~~For the 95th percentile dataset (Table 4; Figures 5–6; Figure 10), the agreement with OPM10 and Ocoarse shifted compared to the full period.~~

~~For OPM10, at AM, NASA_GEOS kept a high correlation close to 0.68 but with large positive bias above +90 μg/m³. MEDIAN showed a slightly lower correlation near 0.56 but lower RMSE and more moderate bias. At BS, NOA_WRF gave the highest correlation around 0.70, but overestimated heavily with bias above +120 μg/m³ and high RMSE. NASA_GEOS gave correlation near 0.62 but again with bias near +100 μg/m³. MEDIAN remained steadier, with correlation between 0.55 and 0.58, lower RMSE than NOA_WRF and NASA_GEOS, and smaller biases. At FKL, NASA_GEOS reached correlation near 0.66 but had high RMSE and bias, while MEDIAN gave correlation above 0.55, the lowest RMSE, and moderate negative bias. DREAMABOL was also reasonable at FKL with correlation close to 0.60 and lower bias than most models. LOTUS and EMA_REG4 again gave the weakest results, with correlations below 0.25 and very high RMSE.~~

~~For Ocoarse, the pattern was similar. NASA_GEOS and NOA_WRF had the highest correlations at AM and BS, but both had very large positive biases. MEDIAN produced moderate correlations but consistently lower RMSE and smaller biases.~~

~~Figure 10 shows that for both O_PM_10 and Ocoarse overestimation dominated during extreme events, especially at BS where NOA_WRF and NASA_GEOS almost entirely overpredicted. MEDIAN reduced this tendency, giving a more even distribution, though its accuracy also dropped compared to the full period.~~

~~In summary, at the 95th percentile NOA_WRF and NASA_GEOS gave the highest correlations with O_{PM10} and Ocoarse but were strongly affected by overestimation. MEDIAN produced more even results with moderate correlation, the lowest RMSE, and smaller biases. DREAMABOL agreed relatively well with O_{PM10} at FKL, while LOTUS and EMA_REG4 gave the weakest results.~~

Table 4: ~~Table 4. Ranked statistics of model performance against PM_{10} (a) and PMcoarse (b) for the 95th percentile of observed concentrations.~~Ranked performance statistics for the 95th percentile (high-dust) subset against observed PM_{10} ($O_{PM_{10}}$) and coarse fraction (O_{coarse}) at Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). Table 3a summarizes results for $O_{PM_{10}}$, and Table 3b for O_{coarse}. Rankings are by performance of R coefficient for each site.~~Rankings are by ….~~

Model	AM (R)	AM_{RMSE} (μg/m³)	AM_{MB} (μg/m³)		BS (R)	BS_{RMSE} (μg/m³)	BS_{MB} (μg/m³)		FKL (R)	FKL_{RMSE} (μg/m³)	FKL_{MB} (μg/m³)
NASA_GEOS	0.66	113.29	41.01	NOA_WRF	0.55	282.49	163.83	NOA_WRF	0.91	66.98	37.45
NOA_WRF	0.53	78.54	25.96	NASA_GEOS	0.49	300.83	83.73	DREAMABOL	0.6	54.98	−15.90
LOTOS-EUROS	0.34	108.52	−20.79	DREAM8b_V2	0.45	253.79	−71.10	NASA_GEOS	0.56	127.04	44.96
NCEP	0.32	78.27	−10.50	NMMB	0.44	315.37	−67.19	MMM	0.55	56.38	−25.42
MMM	0.32	80.27	−35.41	MMM	0.42	252.92	−46.16	DREAM8_MACC	0.53	64.07	−42.74
MACC_ECMWF	0.23	90.59	−43.39	DREAM8_MACC	0.4	263.76	−61.93	NCEP	0.52	67.91	21.92
DREAM8b_V2	0.22	89.8	−42.32	NCEP	0.4	251.04	−86.47	DREAM8b_V2	0.51	63.21	−14.46
DREAM8_MACC	0.2	93.3	−54.97	DREAMABOL	0.39	274.44	−112.28	EMA_REG4	0.48	55.23	−46.97
NMMB	0.18	96.58	−53.53	SILAM	0.39	228.5	11.73	MACC_ECMWF	0.4	88.5	−25.17
DREAMABOL	0.17	79.28	−24.00	MACC_ECMWF	0.34	286.51	−157.67	NMMB	0.27	81.93	−50.52
SILAM	0.1	244.42	25.01	LOTOS-EUROS	0.33	240.86	27.73	SILAM	0.25	76.28	1.48
EMA_REG4	−0.24	117.84	−90.34	EMA_REG4	0.3	447.58	−53.59	LOTOS-EUROS	0.23	13.11	-12.33

(b)

Model	AM (R)	AM_{RMSE} (μg/m³)	AM_{MB} (μg/m³)		BS (R)	BS_{RMSE} (μg/m³)	BS_{MB} (μg/m³)		FKL (R)	FKL_{RMSE} (μg/m³)	FKL_{MB} (μg/m³)
NASA_GEOS	0.63	128.02	68.06	NOA_WRF	0.6	302.59	197.87		–	–	–
LOTOS-EUROS	0.59	54.19	−3.94	NASA_GEOS	0.4	345.94	153.1		–	–	–
NOA_WRF	0.51	87.22	48.4	DREAM8b_V2	0.36	239.84	−9.57		–	–	–
NCEP	0.31	69.67	17.63	NMMB	0.35	324.01	−7.00		–	–	–
MMM	0.31	60.33	−8.59	MMM	0.35	244.69	17		–	–	–
DREAM8_MACC	0.2	68.7	−28.81	DREAM8_MACC	0.34	253.08	2.52		–	–	–
MACC_ECMWF	0.2	70.76	−16.51	DREAMABOL	0.33	226.17	−50.58		–	–	–
DREAM8b_V2	0.18	70.91	−12.69	SILAM	0.33	234.93	43.18		–	–	–
NMMB	0.15	75.86	−25.29	NCEP	0.3	226.39	−22.56		–	–	–
DREAMABOL	0.1	69.74	4.25	MACC_ECMWF	0.28	244.58	−97.75		–	–	–
SILAM	0.05	245.77	56.54	EMA_REG4	0.24	444.59	0.1		–	–	–
EMA_REG4	−0.23	89.6	−63.91	LOTOS-EUROS	0.2	325.55	98.75		–	–	–

(a)

~~Model~~	~~AM (R)~~	~~AM_{RMSE} (μg/m³)~~	~~AM_{MB} (μg/m³)~~	~~BS (R)~~	~~BS_{RMSE} (μg/m³)~~	~~BS_{MB} (μg/m³)~~	~~FKL (R)~~	~~FKL_{RMSE} (μg/m³)~~	~~FKL_{MB} (μg/m³)~~
~~DREAM8_MACC~~	~~0.20~~	~~93.30~~	~~−54.97~~	~~0.40~~	~~263.76~~	~~−61.93~~	~~0.53~~	~~64.07~~	~~−42.74~~
~~DREAM8b_V2~~	~~0.22~~	~~89.80~~	~~−42.32~~	~~0.45~~	~~253.79~~	~~−71.10~~	~~0.51~~	~~63.21~~	~~−14.46~~
~~DREAMABOL~~	~~0.17~~	~~79.28~~	~~−24.00~~	~~0.39~~	~~274.44~~	~~−112.28~~	~~0.60~~	~~54.98~~	~~−15.90~~

EMA_REG4	−0.24	117.84	−90.34	0.30	447.58	−53.59	0.48	55.23	−46.97
LOTUS	0.34	108.52	−20.79	0.33	240.86	27.73	–	–	–
MACC_ECMWF	0.23	90.59	−43.39	0.34	286.51	−157.67	0.40	88.50	−25.17
NASA_GEOS	0.66	113.29	41.01	0.49	300.83	83.73	0.56	127.04	44.96
NCEP	0.32	78.27	−10.50	0.40	251.04	−86.47	0.52	67.91	21.92
NMMB	0.18	96.58	−53.53	0.44	315.37	−67.19	0.27	81.93	−50.52
NOA_WRF	0.53	78.54	25.96	0.55	282.49	163.83	0.91	66.98	37.45
SILAM	0.10	244.42	25.01	0.39	228.50	11.73	0.25	76.28	1.48
MMM	0.32	80.27	−35.41	0.42	252.92	−46.16	0.55	56.38	−25.42

(b)

Model	AM (R)	AM_{RMSE} (µg/m³)	AM_{MB} (µg/m³)	BS (R)	BS_{RMSE} (µg/m³)	BS_{MB} (µg/m³)	FKL (R)	FKL_{RMSE} (µg/m³)	FKL_{MB} (µg/m³)
DREAM8_MACC	0.20	68.70	−28.81	0.34	253.08	2.52	–	–	–
DREAM8b_V2	0.18	70.91	−12.69	0.36	239.84	−9.57	–	–	–
DREAMABOL	0.10	69.74	4.25	0.33	226.17	−50.58	–	–	–
EMA_REG4	−0.23	89.60	−63.91	0.24	444.59	0.10	–	–	–
LOTUS	0.59	54.19	−3.94	0.20	325.55	98.75	–	–	–
MACC_ECMWF	0.20	70.76	−16.51	0.28	244.58	−97.75	–	–	–
NASA_GEOS	0.63	128.02	68.06	0.40	345.94	153.10	–	–	–
NCEP	0.31	69.67	17.63	0.30	226.39	−22.56	–	–	–
NMMB	0.15	75.86	−25.29	0.35	324.01	−7.00	–	–	–
NOA_WRF	0.51	87.22	48.40	0.60	302.59	197.87	–	–	–
SILAM	0.05	245.77	56.54	0.33	234.93	43.18	–	–	–
MMM	0.31	60.33	−8.59	0.35	244.69	17.00	–	–	–

a)

Model	AM_R (R)	AM_RMSE (µg/m³)	AM_MB (µg/m³)	BS_R (R)	BS_RMSE (µg/m³)	BS_MB (µg/m³)	FKL_R (R)	FKL_RMSE (µg/m³)	FKL_MB (µg/m³)
DREAM8_MACC	8 (0.20)	7 (93.30)	11 (-54.97)	6 (0.40)	6 (263.76)	5 (-61.93)	5 (0.53)	5 (64.07)	8 (-42.74)
DREAM8b_V2	7 (0.22)	5 (89.80)	8 (-42.32)	3 (0.45)	5 (253.79)	7 (-71.10)	7 (0.51)	4 (63.21)	2 (-14.46)
DREAMABOL	10 (0.17)	3 (79.28)	3 (-24.00)	8 (0.39)	7 (274.44)	10 (-112.28)	2 (0.60)	1 (54.98)	3 (-15.90)
EMA_REG4	12 (-0.24)	11 (117.84)	12 (-90.34)	12 (0.30)	12 (447.58)	4 (-53.59)	8 (0.48)	2 (55.23)	10 (-46.97)
LOTUS	3 (0.34)	9 (108.52)	2 (-20.79)	11 (0.33)	2 (240.86)	2 (27.73)	-	-	-
MACC_ECMWF	6 (0.23)	6 (90.59)	9 (-43.39)	10 (0.34)	9 (286.51)	11 (-157.67)	9 (0.40)	10 (88.50)	5 (-25.17)
NASA_GEOS	1 (0.66)	10 (113.29)	7 (41.01)	2 (0.49)	10 (300.83)	8 (83.73)	3 (0.56)	11 (127.04)	9 (44.96)
NCEP	4 (0.32)	1 (78.27)	1 (-10.50)	6 (0.40)	3 (251.04)	9 (-86.47)	6 (0.52)	7 (67.91)	4 (21.92)
NMMB	9 (0.18)	8 (96.58)	10 (-53.53)	4 (0.44)	11 (315.37)	6 (-67.19)	10 (0.27)	9 (81.93)	11 (-50.52)
NOA_WRF	2 (0.53)	2 (78.54)	5 (25.96)	1 (0.55)	8 (282.49)	12 (163.83)	1 (0.91)	6 (66.98)	7 (37.45)
SILAM	11 (0.10)	12 (244.42)	4 (25.01)	8 (0.39)	1 (228.50)	1 (11.73)	11 (0.25)	8 (76.28)	1 (1.48)
MEDIAN	4 (0.32)	4 (80.27)	6 (-35.41)	5 (0.42)	4 (252.92)	3 (-46.16)	4 (0.55)	3 (56.38)	6 (-25.42)

(b)

Model	AM_R (R)	AM_RMSE (µg/m³)	AM_MB (µg/m³)	BS_R (R)	BS_RMSE (µg/m³)	BS_MB (µg/m³)	FKL_R (R)	FKL_RMSE (µg/m³)	FKL_MB (µg/m³)
DREAM8_MACC	6 (0.20)	3 (68.70)	8 (-28.81)	6 (0.34)	7 (253.08)	2 (2.52)	-	-	-
DREAM8b_V2	8 (0.18)	7 (70.91)	4 (-12.69)	3 (0.36)	4 (239.84)	4 (-9.57)	-	-	-
DREAMABOL	10 (0.10)	5 (69.74)	2 (4.25)	7 (0.33)	1 (226.17)	8 (-50.58)	-	-	-
EMA_REG4	12 (-0.23)	10 (89.60)	11 (-63.91)	11 (0.24)	12 (444.59)	1 (0.10)	-	-	-
LOTUS	2 (0.59)	1 (54.19)	1 (-3.94)	12 (0.20)	10 (325.55)	10 (98.75)	-	-	-
MACC_ECMWF	6 (0.20)	6 (70.76)	5 (-16.51)	10 (0.28)	5 (244.58)	9 (-97.75)	-	-	-
NASA_GEOS	1 (0.63)	11 (128.02)	12 (68.06)	2 (0.40)	11 (345.94)	11 (153.10)	-	-	-
NCEP	4 (0.31)	4 (69.67)	6 (17.63)	9 (0.30)	2 (226.39)	6 (-22.56)	-	-	-
NMMB	9 (0.15)	8 (75.86)	7 (-25.29)	4 (0.35)	9 (324.01)	3 (-7.00)	-	-	-
NOA_WRF	3 (0.51)	9 (87.22)	9 (48.40)	1 (0.60)	8 (302.59)	12 (197.87)	-	-	-
SILAM	11 (0.05)	12 (245.77)	10 (56.54)	7 (0.33)	3 (234.93)	7 (43.18)	-	-	-
MEDIAN	4 (0.31)	2 (60.33)	3 (-8.59)	4 (0.35)	6 (244.69)	5 (17.00)	-	-	-

Table 4: Performance metrics (R, MB, RMSE) for each dust model during the 95th percentile of observed $O_{PM_{10}}$ and O_{coarse} concentrations at the three monitoring stations: Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). The 95th percentile subset represents extreme dust events. MB and RMSE are reported in µg/m³. Higher R values reflect better temporal agreement, while MB and RMSE quantify the magnitude and direction of forecast error. Results are shown separately for PM_{10} and PM_{coarse}.

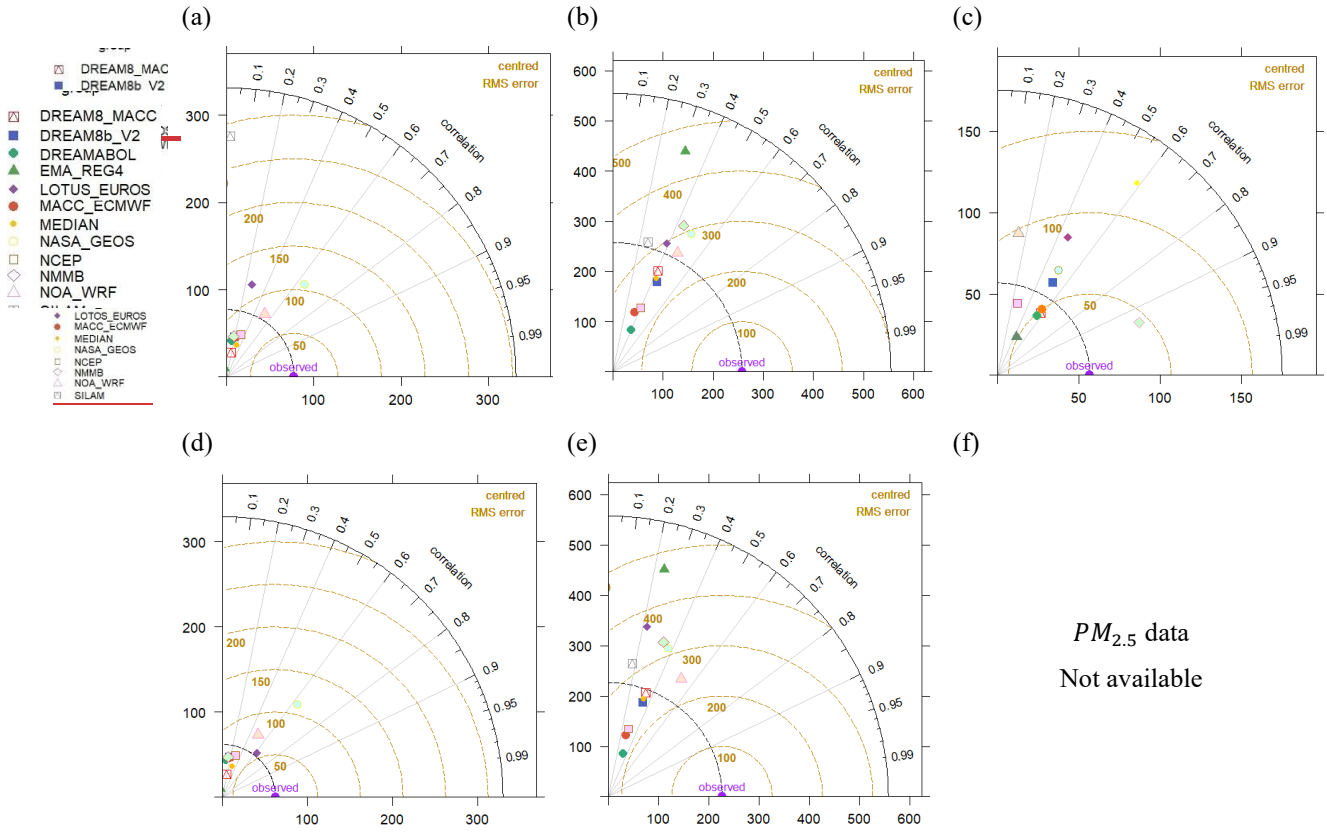


Figure 5: Taylor diagrams showing the performance of 11 operational dust models and the multi-model median (~~MMM~~MEDIAN) during the 95th percentile of observed PM_{10} concentrations at the three monitoring stations: (a and c) Ayia Marina (AM), (b and e) Be'er Sheva (BS) and (c and f) Finokalia (FKL). This analysis reflects the ability of each model to reproduce the timing and magnitude of extreme dust events.

(a) (b)

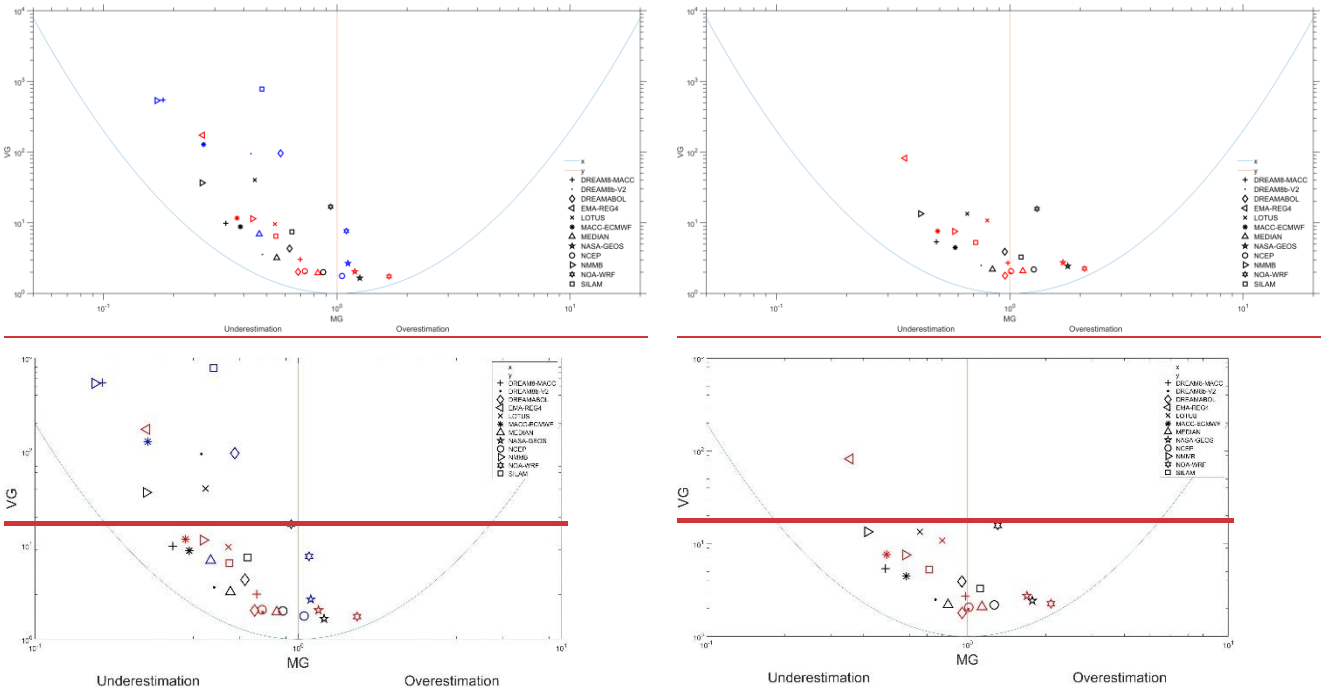


Figure 6: BOOT statistical plots comparing modelled and observed PM_{10} concentrations for the 95th percentile of daily values at Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). a) $M_{dust_{conc}}$ vs $O_{PM_{10}}$ and b) $M_{dust_{conc}}$ vs O_{Coarse}. Models positioned closer to the origin show better performance. Extreme outliers are excluded for readability.

~~4.1.4.~~4.1.3. Results for Dust Days Identified by Achilleos et al. (2020) methodology

~~For the dust-day dataset defined by Achilleos et al. (2019) (Table 5; Figures 7–8; Figure 11), differences between the models and $O_{PM_{10}}$ and O_{coarse} became more pronounced.~~

Figures 6 and 8 and Tables 5a–b summarize model skill on confirmed dust days identified by the Achilleos et al. (2019) methodology. Relative to the full-period results and the high-dust, rankings converge and inter-model spread is reduced, reflecting the stronger event focus of the dust-day selection.

NASA-GEOS and the MMM remain the most consistent performers, sustaining $R \geq 0.60$ at all three sites with $|MB| \leq 40$~~15~~ µg/m³ in most cases and RMSE generally ≤ 75 µg/m³. NOA-WRF retains high event skill, especially at BS and FKL, with R frequently > 0.60 and competitive RMSE, consistent with its improved timing during intense episodes. Systems previously identified as lower-performing (EMA-REG4, ~~LOTUS~~LOTOS-EUROS) continue to exhibit weak correlations ($R < 0.30$) and large errors on dust days, although extreme outliers are fewer than in the 95th percentile analysis.

The dust-day approach and the 95th percentile filter yield broadly consistent rankings for the best (NASA-GEOS, NOA-WRF, MMM) and least-performing (EMA-REG4, ~~LOTUS~~LOTOS-EUROS) models. Notably, MACC-ECMWF and NCEP improve under the dust-day selection, moving up the ranks at one or more sites, indicating that multiparameter event identification (transport, optical/aerosol signals) better aligns model timing and vertical structure with observed peaks. Across models, the dust-day selection reduces bias scatter relative to the 95th percentile subset; for example, at AM, MMM MB improves from −9.84 to 0.66 µg/m³, and DREAM8-MACC from -13.87 to -3.43 µg/m³ when using O_{Coarse} rather than $O_{PM_{10}}$ (see §4.1.1–4.1.2), while RMSE remains in the 20–80 µg/m³ range, implying that mean offsets are mitigated more than amplitude errors.

BS continues to exhibit the largest spread in MB and RMSE (positive MB up to 37 µg/m³, RMSE > 100 µg/m³ for several systems), highlighting persistent challenges in simulating near-source uplift even when analysis is restricted to confirmed dust events. In

contrast, FKL maintains tighter errors (typical RMSE ≈ 20–40 μg/m³) and R ≈ 0.5–0.6, consistent with long-range transport dominance. AM sits between these regimes, with moderate biases and correlations that closely track the all-days ranking.

BOOT plots for dust days (Figure 8) corroborate the metric-based ranking: NASA-GEOS, NOA-WRF, and MMM cluster near the zero-bias axis with reduced dispersion at AM and BS, whereas EMA-REG4, ~~LOTUS~~LOTOS-EUROS, and SILAM appear outside confidence bounds more frequently, indicating persistent magnitude and variability mismatches. Minor layout differences relative to the 95th percentile BOOT do not alter the overall conclusions.

~~Summary.~~ On confirmed dust days, NASA-GEOS and the MMM deliver the most consistent skill (R typically ≥ 0.60, RMSE ≤ ~~75~~45-110 μg m⁻³), NOA-WRF remains event-strong at BS/FKL, and MACC-ECMWF/NCEP show notable gains versus the 95th percentile analysis. The dust-day selection reduces ranking dispersion and bias variability, reinforcing the robustness of the comparative evaluation framework.

~~For $O_{PM_{10}}$, at AM, NASA-GEOS gave the highest correlation (R = 0.65) but with RMSE of 79.13 μg/m³ and bias of +24.70 μg/m³. MEDIAN had correlation 0.56, the lowest RMSE at 53.71 μg/m³, and a negative bias of −14.36 μg/m³. At BS, NOA-WRF gave the highest correlation (R = 0.61) but severely overestimated, with bias +128.66 μg/m³ and RMSE 215.47 μg/m³. NASA-GEOS had correlation 0.59 but also overestimated with bias +75.92 μg/m³. MEDIAN gave correlation 0.51, RMSE 167.23 μg/m³, and smaller bias +7.10 μg/m³. At FKL, NOA-WRF produced the highest correlation (R = 0.68) but bias was again large at +40.68 μg/m³. DREAMABOL also gave strong results at this site with correlation 0.62, RMSE 44.94 μg/m³, and bias −7.56 μg/m³. MEDIAN gave correlation 0.56, the lowest RMSE at 44.88 μg/m³, and bias −12.89 μg/m³.~~

~~For O_{coarse}, NASA-GEOS gave the highest correlation at AM (R = 0.64) but with large positive bias +44.48 μg/m³. NOA-WRF gave the highest correlation at BS (R = 0.64) but the largest bias of all models at +158.60 μg/m³. MEDIAN gave correlation 0.34 at AM, lowest RMSE at 42.89 μg/m³, and near-zero bias +5.29 μg/m³. DREAMABOL again gave good results at FKL with correlation 0.62, RMSE 44.94 μg/m³, and bias −7.56 μg/m³. EMA-REG4, SILAM, and LOTUS consistently produced the lowest correlations, such as EMA-REG4 at AM with R = −0.07, and the highest RMSE values above 190 μg/m³.~~

~~Figures 7–8 show MEDIAN and NASA-GEOS closest to observations, while NOA-WRF appears with high correlation but offset by large bias. BOOT plots confirm that MEDIAN had the tightest distributions, while NOA-WRF and NASA-GEOS were shifted toward overestimation. Figure 11 shows that at BS both NOA-WRF and NASA-GEOS overpredicted strongly for $O_{PM_{10}}$ and O_{coarse}, while MEDIAN gave a more even distribution of under, over, and accurate predictions.~~

~~In summary, during dust days NASA-GEOS and NOA-WRF gave the highest correlations with $O_{PM_{10}}$ and O_{coarse} but consistently overestimated, especially at BS. MEDIAN was more even, with high correlation, the lowest RMSE, and moderate biases. DREAMABOL compared well with OPM10 and O_{coarse} at FKL, while EMA-REG4, SILAM, and LOTUS consistently gave the weakest results.~~ ~~The evaluation of dust days identified by Achilleos et al. (2020) highlights significant variability in model performance across the three sites—Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). The results in this section will not emphasize changes in statistical values with deviations of less than 10%, as these are not considered significant in the overall evaluation of model performance and are excluded for brevity (Table 5).~~

~~At Ayia Marina, correlations are generally low, with only NASA-GEOS (R=0.65) and NOA-WRF (R=0.51) performing well. Most models underestimate dust levels, as reflected in negative mean bias (MB) values, except for NOA-WRF and SILAM, which show overestimations. The RMSE values remain high (> 50 μg/m³), indicating substantial uncertainties during dust events. EMA-REG4 performs particularly poorly, exhibiting a negative correlation (R=−0.07) for this location.~~

~~At Be'er Sheva, model performance is mixed, with only four models—NOA-WRF, MMM, NASA-GEOS, and DREAM8b_V2—achieving correlations above R=0.5. Models like NASA-GEOS, SILAM, and NOA-WRF tend to overestimate PM concentrations,~~

~~while most others exhibit underestimation. High RMSE values, often exceeding 150 μg/m³, reflect the site's proximity to major Middle Eastern dust sources, complicating accurate dust event representation.~~

~~At Finokalia, model correlations are relatively higher, with NOA_WRF (R=0.68) and DREAMABOL (R=0.62) performing best. The station's remote location, less influenced by local anthropogenic activities, allows models to better capture transboundary dust transport. However, overestimations remain common, particularly for NASA-GEOS (MB=43.66 μg/m³).~~

~~Comparison of the two approaches ($O_{PM_{10}}$ vs. O_{coarse}) reveals minimal changes in correlations, but MB often shifts from underestimation to overestimation, particularly at Be'er Sheva and Ayia Marina. This suggests that finer particulate matter in PM_{10} is not adequately captured by models, leading to higher underestimations. The RMSE values remain high for all sites, underscoring the models' uncertainties during extreme dust events.~~

~~The BOOT methodology generally aligns with the MB and RMSE trends but shows notable inconsistencies for some models, particularly those with poor performance. For example, models such as EMA_REG4 and LOTUS display mismatches between their BOOT results and statistical metrics, with BOOT sometimes placing them on the underestimation side even when MB indicates overestimation. Additionally, the BOOT methodology does not fully capture the unique challenges of sites like Be'er Sheva, where proximity to major dust sources introduces higher RMSE values. For example, models such as NASA_GEOS and NOA_WRF perform well in terms of MB and R values but are depicted as having larger scatter on the BOOT graph due to the higher RMSE values typical of this site. Despite these inconsistencies, BOOT provides valuable insights into overall model scatter and bias trends, highlighting areas for further methodological refinement.~~

Table 5: Ranked performance statistics for dust days as defined by Achilleos et al. (2019) against observed PM_{10} ($O_{PM_{10}}$) and coarse fraction (O_{coarse}) at Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). Table 3a summarizes results for $O_{PM_{10}}$, and Table 3b for O_{coarse}. Rankings are by performance of R coefficient -for each site. ~~Rankings are by ….~~

(a)

Model	AM (R)	AM_{RMSE} (μg/m³)	AM_{MB} (μg/m³)		BS (R)	BS_{RMSE} (μg/m³)	BS_{MB} (μg/m³)		FKL (R)	FKL_{RMSE} (μg/m³)	FKL_{MB} (μg/m³)
NASA_GEOS	0.65	79.13	24.7	NOA_WRF	0.61	215.47	128.66	NOA_WRF	0.68	64.9	40.68
NOA_WRF	0.51	79.25	42.39	NASA_GEOS	0.59	212.6	75.92	DREAMABOL	0.62	44.94	−7.56
MMM	0.32	53.71	−14.36	DREAM8b_V2	0.51	165.96	−1.05	NASA_GEOS	0.56	106.02	43.66
MACC_ECMWF	0.26	60.44	−20.19	MMM	0.51	167.23	7.1	MMM	0.56	44.88	−12.89
NCEP	0.25	64.43	11.39	NMMB	0.5	211.68	−37.31	DREAM8b_V2	0.54	51.72	−7.83
DREAM8_MACC	0.22	62.12	−29.31	DREAM8_MACC	0.47	178.55	−13.81	DREAM8_MACC	0.48	51.53	−27.73
DREAM8b_V2	0.22	60.51	−17.55	NCEP	0.46	167.43	−10.07	NCEP	0.46	70.06	31.63
SILAM	0.2	191.44	28.18	MACC_ECMWF	0.42	186.89	−75.36	MACC_ECMWF	0.43	69.81	−17.38
NMMB	0.19	65.52	−28.27	DREAMABOL	0.41	186.52	−22.64	NMMB	0.28	63.45	−33.26
DREAMABOL	0.17	56.11	−4.15	EMA_REG4	0.28	325.89	5.96	EMA_REG4	0.24	50.77	−29.62
LOTOS-EUROS	0.15	81.04	−7.20	SILAM	0.26	189.39	32.42	SILAM	0.19	72.42	16.3
EMA_REG4	−0.07	69.43	−48.03	LOTOS-EUROS	0.08	258.6	27.86	LOTOS-EUROS	–	–	–

(b)

Model	AM (R)	AM_{RMSE} (μg/m³)	AM_{MB} (μg/m³)		BS (R)	BS_{RMSE} (μg/m³)	BS_{MB} (μg/m³)		FKL (R)	FKL_{RMSE} (μg/m³)	FKL_{MB} (μg/m³)
NASA_GEOS	0.64	89.58	44.48	NOA_WRF	0.64	237.19	158.6				
NOA_WRF	0.52	89.31	60.03	NASA_GEOS	0.52	241.7	118.21				
MMM	0.34	42.89	5.29	MMM	0.47	166.29	47.51				
MACC_ECMWF	0.26	49.57	−0.43	DREAM8b_V2	0.44	162.37	39.91				
NCEP	0.26	64.19	31.01	DREAM8_MACC	0.43	172.53	26.37				
DREAM8_MACC	0.23	44.95	−9.77	NMMB	0.43	216.35	2.08				
DREAM8b_V2	0.22	50.04	1.65	NCEP	0.39	157.88	29.14				
NMMB	0.2	52.03	−8.72	DREAMABOL	0.37	161.09	14.82				
SILAM	0.19	195.98	45.78	MACC_ECMWF	0.37	159.61	−35.71				
DREAMABOL	0.18	51.11	14.04	EMA_REG4	0.25	333.95	43.98				

LOTOS-EUROS	0.18	77.39	11.13	SILAM	0.23	198.82	58.12				
EMA_REG4	−0.07	50.5	−29.63	LOTOS-EUROS	0.1	253.05	50.9				

Table 5. Ranked statistics of model performance against PM_{10} (a) and PM_{coarse} (b) for dust days as defined by Achilleos et al. (2019).

(a)

Model	AM_R (R)	AM_RMSE (μg/m³)	AM_MB (μg/m³)	BS_R (R)	BS_RMSE (μg/m³)	BS_MB (μg/m³)	FKL_R (R)	FKL_RMSE (μg/m³)	FKL_MB (μg/m³)
DREAM8_MACC	6 (0.22)	5 (62.12)	10 (-29.31)	6 (0.47)	4 (178.55)	5 (-13.81)	6 (0.48)	4 (51.53)	6 (-27.73)
DREAM8b_V2	6 (0.22)	4 (60.51)	5 (-17.55)	3 (0.51)	1 (165.96)	1 (-1.05)	5 (0.54)	5 (51.72)	2 (-7.83)
DREAMABOL	10 (0.17)	2 (56.11)	1 (-4.15)	9 (0.41)	5 (186.52)	6 (-22.64)	2 (0.62)	2 (44.94)	1 (-7.56)
EMA_REG4	12 (-0.07)	8 (69.43)	12 (-48.03)	10 (0.28)	12 (325.89)	2 (5.96)	10 (0.24)	3 (50.77)	7 (-29.62)
LOTUS	11 (0.15)	11 (81.04)	2 (-7.20)	12 (0.08)	11 (258.60)	7 (27.86)	-	-	-
MACC_ECMWF	4 (0.26)	3 (60.44)	6 (-20.19)	8 (0.42)	6 (186.89)	10 (-75.36)	8 (0.43)	8 (69.81)	5 (-17.38)
NASA_GEOS	1 (0.65)	9 (79.13)	7 (24.70)	2 (0.59)	9 (212.60)	11 (75.92)	3 (0.56)	11 (106.02)	11 (43.66)
NCEP	5 (0.25)	6 (64.43)	3 (11.39)	7 (0.46)	3 (167.43)	4 (-10.07)	7 (0.46)	9 (70.06)	8 (31.63)
NMMB	9 (0.19)	7 (65.52)	9 (-28.27)	5 (0.50)	8 (211.68)	9 (-37.31)	9 (0.28)	6 (63.45)	9 (-33.26)
NOA_WRF	2 (0.51)	10 (79.25)	11 (42.39)	1 (0.61)	10 (215.47)	12 (128.66)	1 (0.68)	7 (64.90)	10 (40.68)
SILAM	8 (0.20)	12 (191.44)	8 (28.18)	11 (0.26)	7 (189.39)	8 (32.42)	11 (0.19)	10 (72.42)	4 (16.30)
MEDIAN	3 (0.32)	1 (53.71)	4 (-14.36)	3 (0.51)	2 (167.23)	3 (7.10)	3 (0.56)	1 (44.88)	3 (-12.89)

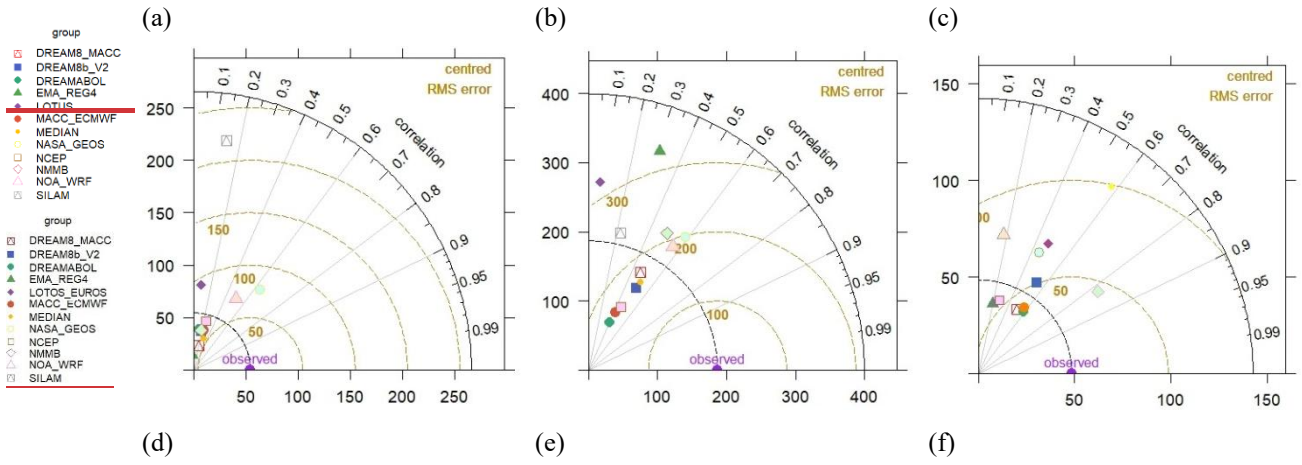
(b)

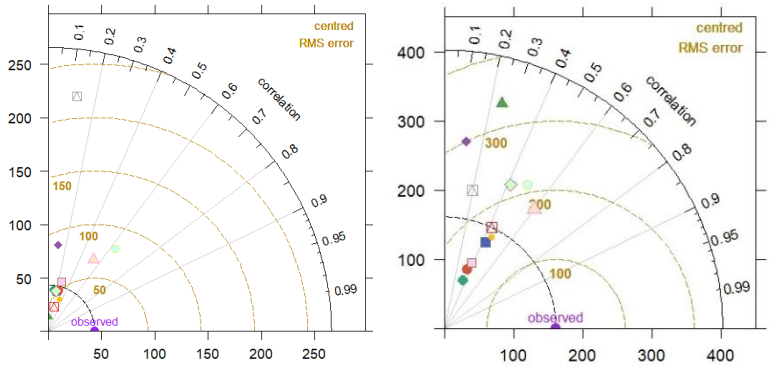
Model	AM_R (R)	AM_RMSE (μg/m³)	AM_MB (μg/m³)	BS_R (R)	BS_RMSE (μg/m³)	BS_MB (μg/m³)	FKL_R (R)	FKL_RMSE (μg/m³)	FKL_MB (μg/m³)
DREAM8_MACC	6 (0.23)	2 (44.95)	5 (-9.77)	5 (0.43)	6 (172.53)	3 (26.37)	-	-	-
DREAM8b_V2	7 (0.22)	4 (50.04)	2 (1.65)	4 (0.44)	4 (162.37)	6 (39.91)	-	-	-
DREAMABOL	10 (0.18)	6 (51.11)	7 (14.04)	8 (0.37)	3 (161.09)	2 (14.82)	-	-	-
EMA_REG4	12 (-0.07)	5 (50.50)	8 (-29.63)	10 (0.25)	12 (333.95)	7 (43.98)	-	-	-
LOTUS	10 (0.18)	9 (77.39)	6 (11.13)	12 (0.10)	11 (253.05)	9 (50.90)	-	-	-
MACC_ECMWF	4 (0.26)	3 (49.57)	1 (-0.43)	8 (0.37)	2 (159.61)	5 (-35.71)	-	-	-
NASA_GEOS	1 (0.64)	11 (89.58)	10 (44.48)	2 (0.52)	10 (241.70)	11 (118.21)	-	-	-
NCEP	4 (0.26)	8 (64.19)	9 (31.01)	7 (0.39)	1 (157.88)	4 (29.14)	-	-	-
NMMB	8 (0.20)	7 (52.03)	4 (-8.72)	5 (0.43)	8 (216.35)	1 (2.08)	-	-	-
NOA_WRF	2 (0.52)	10 (89.31)	12 (60.03)	1 (0.64)	9 (237.19)	12 (158.60)	-	-	-
SILAM	9 (0.19)	12 (195.98)	11 (45.78)	11 (0.23)	7 (198.82)	10 (58.12)	-	-	-
MEDIAN	3 (0.34)	1 (42.89)	3 (5.29)	3 (0.47)	5 (166.29)	8 (47.51)	-	-	-

Table 5: Statistical performance metrics (correlation coefficient R, mean bias MB, and Root Mean Square Error RMSE) for modelled dust concentrations compared to observed $O_{PM_{10}}$ and O_{coarse} values at three background stations (Ayia Marina – AM, Be'er Sheva – BS, and Finokalia – FKL) over for the Achilleos et al. (2019) approach.

Model	Station	$O_{PM_{10}}$			O_{coarse}		
		R	MB (μg/m³)	RMSE (μg/m³)	R	MB (μg/m³)	RMSE (μg/m³)
DREAM8_MACC	AM	0.22	-29.31	62.12	0.23	-9.77	44.95
	BS	0.47	-13.81	178.55	0.43	26.37	172.53
	FKL	0.48	-27.73	51.53	-	-	-
DREAM8b_V2	AM	0.22	-17.55	60.51	0.22	1.65	50.04
	BS	0.51	-1.05	165.96	0.44	39.91	162.37
	FKL	0.54	-7.83	51.72	-	-	-
DREAMABOL	AM	0.17	-4.15	56.11	0.18	14.04	51.11
	BS	0.41	-22.64	186.52	0.37	14.82	161.09
	FKL	0.62	-7.56	44.94	-	-	-

EMA_REG4	AM	-0.07	-48.03	69.43	-0.07	-29.63	50.50
	BS	0.28	5.96	325.89	0.25	43.98	333.95
	FKL	0.24	-29.62	50.77	-	-	-
LOTUS	AM	0.15	-7.20	81.04	0.18	11.13	77.39
	BS	0.08	27.86	258.60	0.10	50.90	253.05
	FKL	-	-	-	-	-	-
MACC_ECMWF	AM	0.26	-20.19	60.44	0.26	-0.43	49.57
	BS	0.42	-75.36	186.89	0.37	-35.71	159.61
	FKL	0.43	-17.38	69.81	-	-	-
MMM	AM	0.32	-14.36	53.71	0.34	5.29	42.89
	BS	0.51	7.10	167.23	0.47	47.51	166.29
	FKL	0.56	-12.89	44.88	-	-	-
NASA_GEOS	AM	0.65	24.70	79.13	0.64	44.48	89.58
	BS	0.59	75.92	212.60	0.52	118.21	241.70
	FKL	0.56	43.66	106.02	-	-	-
NCEP	AM	0.25	-11.39	64.43	0.26	31.01	64.19
	BS	0.46	-10.07	167.43	0.39	29.14	157.88
	FKL	0.46	31.63	70.06	-	-	-
NMMB	AM	0.19	-28.27	65.52	0.20	-8.72	52.03
	BS	0.50	-37.31	211.68	0.43	2.08	216.35
	FKL	0.28	-33.26	63.45	-	-	-
NOA_WRF	AM	0.51	42.39	79.25	0.52	60.03	89.31
	BS	0.61	128.66	215.47	0.64	158.60	237.19
	FKL	0.68	40.68	64.90	-	-	-
SILAM	AM	0.20	28.18	191.44	0.19	45.78	195.98
	BS	0.26	32.42	189.39	0.23	58.12	198.82
	FKL	0.19	16.30	72.42	-	-	-





$PM_{2.5}$ data

Not available

Figure 7: Taylor diagrams for daily PM_{10} during high-dust conditions (95th percentile of observed concentrations) at the three monitoring sites. Panels show model performance against observations at (a) Ayia Marina (AM), (b) Be'er Sheva (BS), and (c) Finokalia (FKL). The azimuth indicates the correlation coefficient (R), the radial distance the standard deviation (SD), and the concentric arcs the centered RMSE. Symbols denote the 11 operational models; the multi-model median (MMM) is highlighted. The observational reference is located at R=1 and SD=1. ~~Taylor diagrams showing the performance of 11 operational dust models and the multi-model median (MMMMEDIAN) during identified dust storm days at the three monitoring stations: (a and c) Ayia Marina (AM), (b and e) Be'er Sheva (BS) and (c and f) Finokalia (FKL).~~

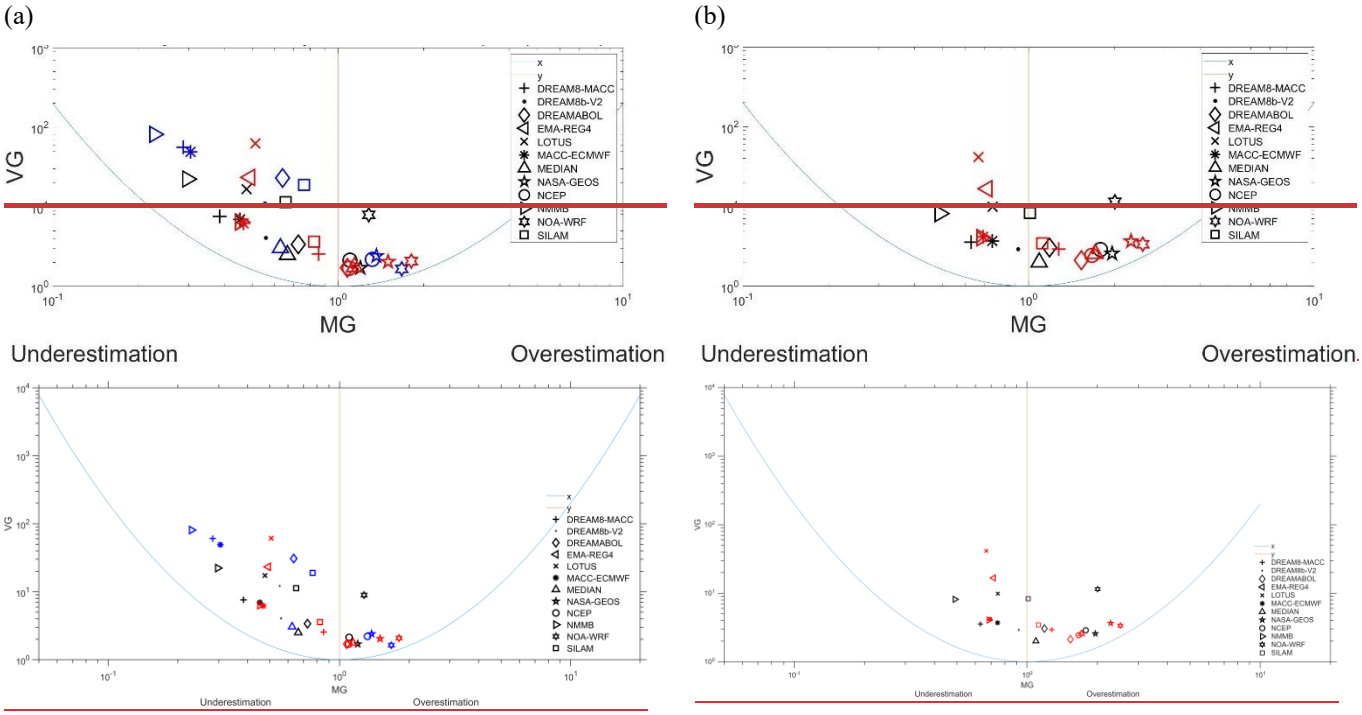


Figure 8: BOOT statistical plots comparing modelled and observed PM_{10} concentrations for the Achilleos et al. (2019) approach at Ayia Marina (AM), Be'er Sheva (BS), and Finokalia (FKL). a) $M_{dust_{conc}}$ vs $O_{PM_{10}}$ and b) $M_{dust_{conc}}$ vs O_{Coarse}. Models positioned closer to the origin show better performance. Extreme outliers are excluded for readability.

~~Summarizing the results based on the Achilleos et al. (2019) approach and site-specific analysis, Ayia Marina shows low overall correlation values for most models, with only NASA-GEOS (R = 0.66) and NOA-WRF (R = 0.53) performing well. The remaining models fail to capture dominant dust patterns, indicating insufficient representation of key dust sources for this site. EMA-REG4 displays negative correlation (R = −0.24), clearly failing to simulate observed conditions. Most models tend to underestimate major dust events, with only a few (e.g., NASA-GEOS, SILAM) producing slight overestimations. In two of the three sites, models with higher correlation also exhibit stronger overall performance.~~

~~At Be'er Sheva, model performance is generally poor. Only NOA-WRF (R = 0.62), MMM MEDIAN (R = 0.58), NASA-GEOS (R = 0.65), and DREAM8b_V2 (R = 0.45) exceed R = 0.5. Underestimation dominates, though models like NASA-GEOS, SILAM, and NOA-WRF show positive mean bias values (e.g., MB = 197.87 µg/m³ for NOA-WRF). The site's proximity to Middle Eastern dust sources likely contributes to this variability, as reflected in the significantly higher RMSE values compared to Ayia Marina and Finokalia. At Finokalia, most models perform comparatively better. NOA-WRF (R = 0.91) and DREAMABOL (R = 0.60) achieve the highest correlations. The site's limited anthropogenic influence and exposure to transboundary dust enhance model agreement with observations.~~

~~Comparing the two evaluation approaches, $O_{PM_{10}}$ and O_{coarse}, reveals minor differences in R coefficient, but notable changes in MB. In several cases, MB shifts from negative to positive or vice versa. A general trend is observed: underestimating models reduce their negative bias, and slightly overestimating models exhibit larger positive bias under the O_{coarse} approach. This suggests that finer particulate matter in $O_{PM_{10}}$ is not fully captured, leading to stronger underestimations, particularly for models that do not resolve $PM_{2.5}$ well.~~

~~RMSE remains consistently high across all sites, confirming persistent model uncertainty during intense dust events. The BOOT evaluation supports the MB trends but shows discrepancies for low-performing models, further emphasizing the limitations in model reliability under extreme conditions.~~

4.2. Categorical statistics

~~The following results present the contingency tables in the form of heatmaps, depicting the daily agreement of modelled and observed values at each site. The results are categorized into three scenarios: overestimation, underestimation and accurate prediction (hit; within 1 SD of the ground observations) for each approach and dataset used. The two approaches include O_{coarse} and $O_{PM_{10}}$, focusing on how well the models capture different PM fractions at each site.~~The following results present contingency table analyses in the form of daily heatmaps, illustrating the agreement between modelled and observed PM concentrations at each site. For each dataset and approach, model performance is categorized into three outcomes: overestimation, underestimation, and accurate prediction (hits, defined as model values within ±1 standard deviation of ground observations). Two modelling approaches are evaluated—O_{coarse} and $O_{PM_{10}}$ to assess how effectively each model captures different particulate matter fractions and their day-to-day variability across sites.

4.2.1. Ayia Marina (AM), Cyprus

Figure 9: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the entire study period. Color intensity shows the density of paired values, indicating the level of agreement between modeled dust and observed particulate concentrations.

~~Figure 9~~a and b show the heatmaps for the Ayia Marina site, comparing model performance for O_{coarse} and $O_{PM_{10}}$ over the entire study period. All models achieve over 80% hits for both approaches, demonstrating their ability to capture the overall dust conditions at this site. The DREAM8-MACC model stands out with the highest hit rat`e at 97%, indicating excellent performance in predicting both coarse and total PM concentrations. On the other hand, NOA-WRF records the lowest hit rate (81%), reflecting a tendency to overestimate the dust levels in this region, as confirmed by the model's consistent overestimation across multiple metrics in prior analyses. Underestimation remains minimal for Ayia Marina in both approaches, indicating that most models are

capturing or over-predicting dust events rather than missing them. This is critical for public health warnings, as underestimation would lead to insufficient precautionary measures.

For the 95th percentile of ground observations (Figure 10: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the 95th percentile of observed data. Color intensity indicates the density of paired values, highlighting model–observation agreement during extreme dust events

~~Figure 10~~a and b), AM shows mixed results between hits and overestimations in the O_{coarse} approach. NOA-WRF and NASA-GEOS display the highest overestimations, consistent with their tendency to over-predict extreme dust events, which has been observed across multiple metrics. EMA-REG4, on the other hand, records the fewest overestimations and is more prone to underestimating, continuing its underperformance noted in earlier sections. The overall hit percentages remain high for most models, but overestimations dominate during peak dust events. In the $O_{PM_{10}}$ approach, the results are similar, with NOA-WRF and NASA-GEOS again leading in overestimations, but no strong trends emerge beyond these individual cases. EMA-REG4 again shows high underestimation, making it the weakest performer during these peak dust events.

For the specific dates examined, following the third approach of Subsection 3.3, the O_{coarse} approach shows (Figure 10: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the 95th percentile of observed data. Color intensity indicates the density of paired values, highlighting model–observation agreement during extreme dust events

~~Figure 10~~a) that most models perform well, capturing ground concentrations within one standard deviation. EMA-REG4 has the highest hit percentage, which is notable given its underperformance in other scenarios. DREAM8-MACC also performs well, while NOA-WRF shows the highest overestimation rate, continuing the trend seen in previous sections. Underestimations remain low, indicating that most models effectively predict peak dust levels.

4.2.2. Be'er Sheva (BS), Israel

The performance at Be'er Sheva, shown in Figure 9: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the entire study period. Color intensity shows the density of paired values, indicating the level of agreement between modeled dust and observed particulate concentrations.

~~Figure 9~~a and b, is more varied compared to AM. For the O_{coarse} approach, the models exhibit a broader range of results. MACC-ECMWF and NMMB models show the highest hit percentages, indicating strong predictive capabilities for coarse particles. In contrast, DREAM8b-V2 has the lowest hit rate, suggesting a struggle to accurately capture dust levels at this location, particularly for coarse particles. In the $O_{PM_{10}}$ approach, the performance remains largely the same, with MACC-ECMWF and NMMB continuing to outperform other models. However, DREAM8b-V2 again records the lowest hit rate, reinforcing its overall weaker performance at this site, which was also observed in the earlier analyses of correlation and bias metrics. Overestimation is more frequent than underestimation, particularly for NOA-WRF, which tends to significantly overestimate PM concentrations, leading to potential overreactions in dust event management.

For the 95th percentile of ground observations, overestimations dominate the results for both approaches (O_{coarse} and $O_{PM_{10}}$), as depicted in Figure 10: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse

fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the 95th percentile of observed data. Color intensity indicates the density of paired values, highlighting model–observation agreement during extreme dust events

~~Figure 10~~a and b. This site's proximity to major dust sources in the Middle East may contribute to the over-prediction, as models struggle to account for rapid dust inflow. NOA-WRF shows the highest overestimation, with no underestimations recorded for the O_{coarse} approach, reinforcing the model's consistent bias towards overestimating dust levels at this site. EMA-REG4, which previously showed underestimations at other sites, also demonstrates high underestimations for $O_{PM_{10}}$, indicating that it struggles to capture the dust dynamics in Be'er Sheva.

For the specific dates examined, at Be'er Sheva, overestimations dominate again in the O_{coarse} approach, with NOA-WRF and NASA-GEOS showing the highest overestimation percentages. EMA-REG4, typically prone to underestimations, continues to show a relatively balanced performance between hits and underestimations. In the $O_{PM_{10}}$ approach (Figure 11b), the results are consistent, with high overestimation rates, reinforcing the models' struggle to accurately predict dust levels in this challenging region.

4.2.3. Finokalia (FKL), Greece

At Finokalia, ~~the~~ model~~s~~ ~~generally~~ performance is consistently strong ~~well~~, with all ~~showing~~ systems achieving hit percentages above 88%~~,~~ ~~as illustrated~~(~~in~~ Figure 9: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the entire study period. Color intensity shows the density of paired values, indicating the level of agreement between modeled dust and observed particulate concentrations.

~~Figure 9~~). The ~~LOTUS~~LOTOS-EUROS model performs ~~exceptionally well in both approaches~~best, ~~achieving~~ reaching a 100% hit rate for O_{coarse}, indicating that it accurately captures dust conditions at this site. This high performance contrasts with ~~LOTUS~~LOTOS-EUROS's weaker results at other sites, like Be'er Sheva, highlighting its region-specific strengths. Conversely, NCEP records the lowest hit percentage, suggesting challenges in predicting both coarse and total PM at Finokalia. Finokalia, with its more consistent results, supports the idea that this site, being less influenced by anthropogenic sources, allows models to focus more on transboundary dust transport, leading to higher accuracy across most models. This observation aligns with the results from previous sections, where high correlations and lower biases were recorded for many models at this site.

For this area, there are no trends using the 95th percentile of ground observations, in either approach, with most models showing a balance between hits, overestimations, and underestimations (Figure 10: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the 95th percentile of observed data. Color intensity indicates the density of paired values, highlighting model–observation agreement during extreme dust events

~~Figure 10~~). EMA-REG4 and NMMB record the highest underestimations, while NOA-WRF shows the highest overestimations, consistent with its overall behaviour across the sites. The hit percentages remain relatively high, but the variability indicates that extreme dust events in Finokalia are more challenging for models to predict accurately, likely due to the site's geographical characteristics.

For the specific-dates approach~~Finally, for specific dates approach~~, ~~and~~ using $O_{PM_{10}}$ data, ~~results show~~ (Figure 11: Heatmaps of modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed

PM10 ($O_{PM_{10}}$) for dust days (Achilleos et al., 2019). Color intensity shows the frequency of paired values, illustrating model performance under dust-dominated conditions.

~~Figure 11~~b), ~~that~~ most models exceed ~~achieve more than~~ 50% hit rates, with DREAMABOL performing best ~~achieving the highest hit percentage~~. NOA-WRF remains the only model to exceed 80% overestimation, reinforcing its bias toward higher PM_{10} levels during dust events~~tendency to over-predict dust levels at this site~~. Overall, Finokalia exhibits robust model agreement, but continued overestimation of peak magnitudes underscores the need for improved representation of intense long-range dust transport episodes~~These results align with earlier observations, suggesting that while most models perform well at Finokalia, there is still room for improvement in accurately capturing peak dust events~~.

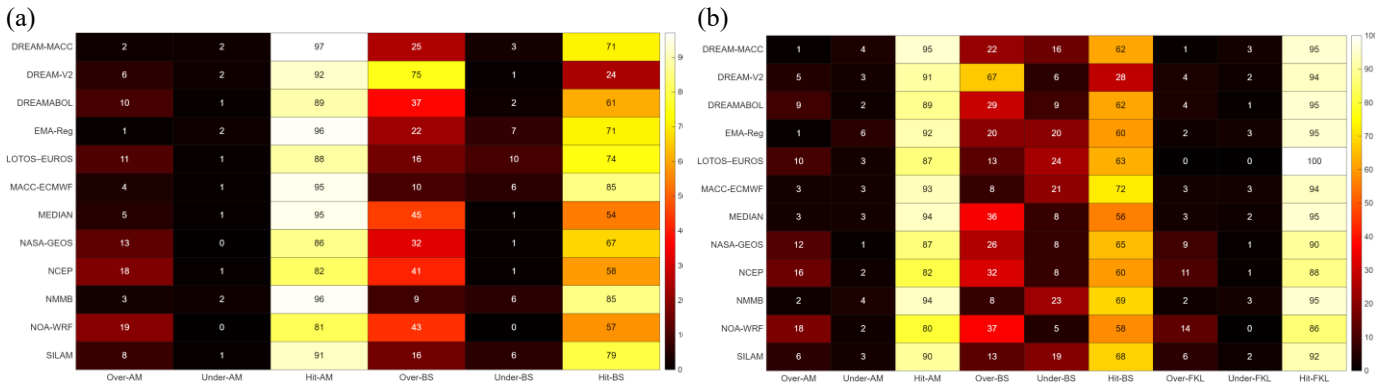


Figure 9:~~-~~ ~~-~~Heatmaps of ~~modeled~~modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the entire study period. Color intensity shows the density of paired values, indicating the level of agreement between modeled dust and observed particulate concentrations.

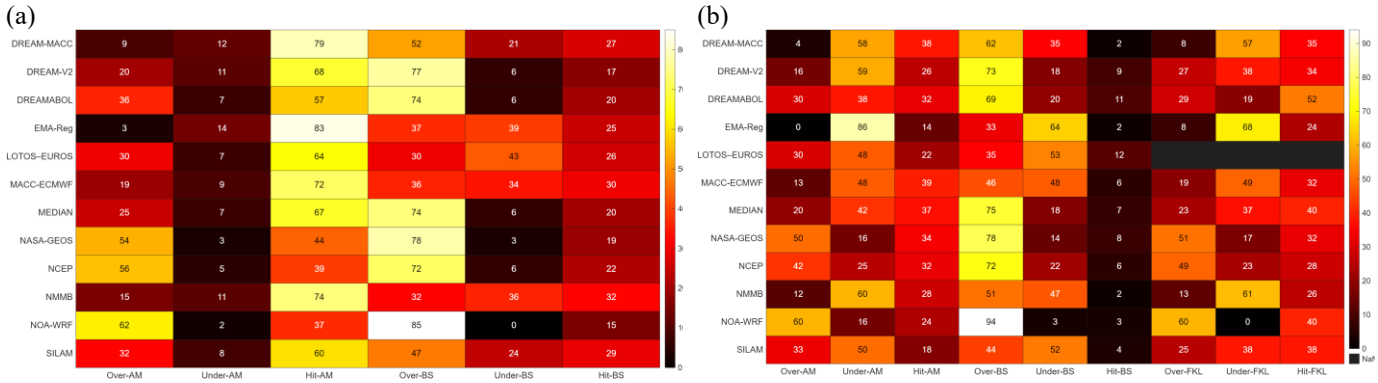


Figure 10:~~-~~ Heatmaps of ~~modeled~~modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM_{10} ($O_{PM_{10}}$) for the 95th ~~-~~percentile of observed data. Color intensity indicates the density of paired values, highlighting model–observation agreement during extreme dust events

(a) (b)

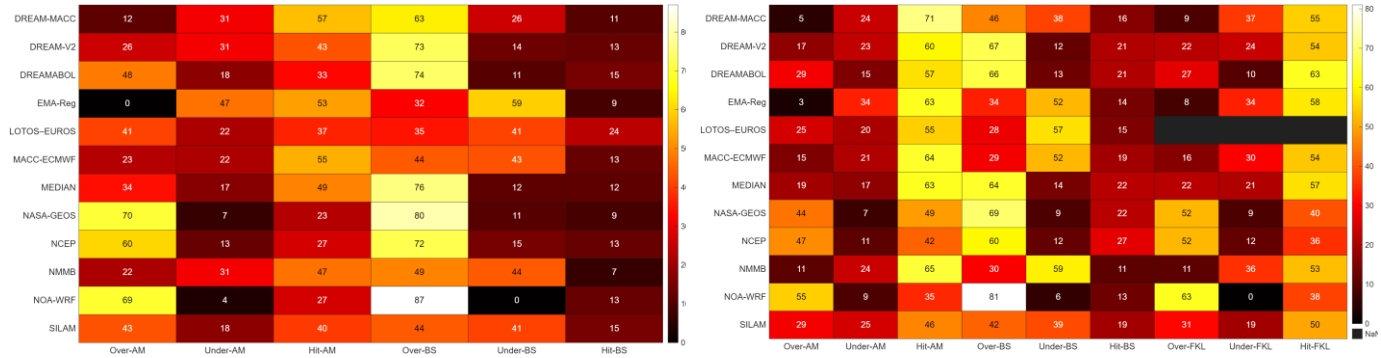


Figure 11: ~~.~~ Heatmaps of ~~modeled~~modelled dust concentrations ($M_{dust_{conc}}$) against (a) observed coarse fraction (O_{coarse}) and (b) observed PM10 ($O_{PM_{10}}$) for dust days (Achilleos et al., 2019). Color intensity shows the frequency of paired values, illustrating model performance under dust-dominated conditions.

~~**Figure 11: Heatmap of the scenarios (a) $M_{dust_{conc}}$ vs O_{coarse} and (b) $M_{dust_{conc}}$ vs $O_{PM_{10}}$ for specific dust days.**~~

Overall, no model demonstrates exceptional accuracy across all sites, approaches and datasets. A high hit rate of models accurately predicting observations is observed only in the first dataset, which assesses performance over the entire study period. Interestingly, models like EMA-REG4 and NMMB, which consistently perform poorly in other evaluation methods, show significantly better results when evaluated using contingency table methods.

5. Evaluation of Configuration Settings in the performance of Operational Dust Forecasting Models

Differences in model configuration lead to measurable variations in performance across EMR sites. Key configuration elements shaping model performance include the representation of the near-surface vertical structure (including first-layer height), the use of aerosol and meteorological data assimilation and the horizontal resolution; their relative influence varies by site and evaluation subset. Over the full period, systems that combine fine near-surface vertical structure with data assimilation, most notably NASA-GEOS, achieve higher correlations and balanced errors across sites (e.g., R = 0.71 at AM, 0.65 at BS, and 0.64 at FKL for PM_{10}), with $|MB| \leq 15$ μg/m³ and RMSE typically ≤ 75 μg/m³ (Table 3a–b; Figures 3–4). The MMM similarly exhibits near-leading, stable performance by reducing the influence of outliers at each timestamp and delivering top-rank RMSE with moderate-to-high correlations.

The influence of vertical structure is most evident in the representation of surface mixing and diurnal gradients. Finer vertical discretization and lower first-layer heights consistently improve correlation and reduce RMSE across sites. Conversely, coarser vertical structures under-capture near-surface variability at FKL and tend to amplify positive bias at BS, shifting MB away from zero. These effects are substantiated in the updated Table 2, which reports first-layer heights (as provided by modeling teams), clarifying cross-system differences that align with the observed rankings in Table 3.

Horizontal resolution primarily affects the timing and amplitude of intense dust episodes rather than average performance across the full period. NOA-WRF illustrates this pattern: it achieves strong event-level performance (e.g., R = 0.91 at FKL in the 95th percentile subset) and maintains competitive full-period skill at BS (R = 0.62). However, it also overestimates near-source concentrations, with MB ≈ +10 to +40 μg/m³ and RMSE in the 70–110 μg/m³ range. At sites where long-range transport dominates, such as Finokalia and Ayia Marina, RMSE values are lower (≈ 20–40 μg/m³ at FKL and 25–35 μg/m³ at AM), and finer grid resolution provides limited additional benefit once transport phasing is adequately captured.

The application of data assimilation systematically improves transport phasing and magnitude, raising correlations and reducing biases across both full-period and event-focused subsets. NASA-GEOS, which assimilates relevant meteorological and aerosol

fields, consistently maintains $R \geq 0.60$ across all sites and subsets, with MB near zero and RMSE ≤ 75 μg/m³. Other data assimilation -enabled models (e.g., MACC-ECMWF, NCEP) show ranking gains under confirmed dust-day conditions relative to the 95th percentile subset, suggesting that multi-parameter event identification enhances the benefit of data assimilation by better aligning vertical structure and timing with observations.

Additional factors such as meteorological forcing type (e.g., analysis vs forecast) and source emission parameterizations further influence site-specific skill. At BS, close to major dust sources, models exhibit the largest variance and error spread (MB up to +37 μg/m³; RMSE > 100 μg/m³ in several systems), reflecting sensitivity to uplift intensity and boundary inflow. In contrast, AM and FKL (representing regional background conditions) display more systematic underestimation (MB ≈ −5 to −15 μg/m³) and lower RMSE values (25–35 μg/m³ at AM; 20–40 μg/m³ at FKL), consistent with reduced influence from local sources and more predictable transport pathways.

Within this context, the MMM provides robust central performance by computing the median across harmonized model outputs at each timestamp and comparing the resulting time series to observations. Over the full period, MMM achieves $R \geq 0.60$ at one or more sites, with low RMSE and $|MB| \leq 15$ μg/m³. Substitution of O_{coarse} for PM_{10} generally reduces mean biases without significantly altering amplitude errors: at AM, MMM MB improves from −9.84 to +0.66 μg/m³, and DREAM8-MACC from −13.87 to −3.43 μg/m³, while RMSE remains within the range of ~20–80 μg/m³. This suggests that coarse-fraction filtering corrects offset but does not resolve the full amplitude spread during dust events.

In summary, the most influential configuration elements are:

(i) vertical resolution near the surface, including first-layer height, and

(ii) the presence of data assimilation.

Horizontal resolution contributes less to full-period averages but enhances event timing and intensity during extremes. These relationships explain the consistent top-tier performance of NASA-GEOS, the stability of the MMM, and the event-specific strengths (and biases) of NOA-WRF. From an operational perspective, MMM-guided forecasts, supported by data assimilation-enabled systems with fine vertical resolution and routine evaluation against both $O_{PM_{10}}$ and O_{coarse}, offer a practical strategy to reduce bias, limit error growth during dust intrusions, and enhance day-to-day forecast reliability across the EMR.

~~The evaluation across the entire period (Table 3), the 95th percentile (Table 4), and dust days (Table 5) for $O_{PM_{10}}$ and O_{coarse}, supported by Figures 3–11, shows clear links between model configuration (Tables 1–2) and forecasting results.~~

~~Vertical resolution influenced agreement with observations. NASA-GEOS, with many vertical levels and assimilation, achieved of the highest correlations for $O_{PM_{10}}$ in all datasets: 0.71, 0.65, and 0.64 for the entire period (AM, BS, FKL), 0.66, 0.49, and 0.56 for the 95th percentile, and about 0.68 at AM on dust days. For O_{coarse}, correlations were lower, typically 0.59–0.62 during the entire period, and BOOT plots showed more scatter. LOTUS, with fewer vertical levels and limited coverage, gave weak values throughout, with $O_{PM_{10}}$ correlations as low as 0.15 on dust days at AM and O_{coarse} correlation of 0.18. Figures 3, 5, and 7 place NASA-GEOS close to the observation reference, while LOTUS consistently lies farther away.~~

~~First-layer height affected site-specific behaviour. NCEP compared well with $O_{PM_{10}}$ at FKL ($R = 0.62$ for the entire period, 0.52 for the 95th percentile, and 0.54 on dust days) but was weaker at AM ($R \approx 0.27$ entire period, 0.32 for the 95th). MACC-ECMWF also started low near the surface and achieved moderate $O_{PM_{10}}$ correlation around 0.5 for the entire period, but at the 95th percentile dropped to 0.40 or below. Both models showed broader dispersion when compared with O_{coarse}, as seen in Figures 9–11.~~

~~Horizontal resolution captured variability but also introduced systematic biases. NOA-WRF gave high $O_{PM_{10}}$ correlation at FKL (0.58 entire period, 0.91 for the 95th percentile, 0.69 dust days). At BS, however, $O_{PM_{10}}$ correlation was 0.62 for the entire period but with strong positive bias, which increased further in the 95th percentile. For Ocoarse, correlations were lower, typically between~~

0.50 and 0.60. DREAMABOL also matched $O_{PM_{10}}$ well at FKL (0.61–0.63 entire period, ~0.60 dust days) but less so at BS. Figures 9–11 confirm NOA-WRF's tendency to overpredict $O_{PM_{10}}$, while O_{coarse} results were less biased but more dispersed.

Data assimilation and meteorological drivers shaped model differences. NASA-GEOS, driven by GEOS-5, consistently gave the highest $O_{PM_{10}}$ correlations across datasets but showed growing biases at BS for the 95th percentile and dust days. MACC-ECMWF, despite assimilation, showed weaker agreement with $O_{PM_{10}}$ (≤0.40 at the 95th percentile, with large negative biases) and poor representation of O_{coarse}. DREAM8_MACC and DREAM8b_V2 produced intermediate values, for example O_{coarse} at BS on dust days with correlation around 0.52 and biases between −14 and −31 μg/m³. EMA-REG4 and SILAM performed weakly across all datasets, with very low correlations and large errors.

The multi-model median (MEDIAN) was consistently balanced. For $O_{PM_{10}}$, it often achieved the lowest RMSE, with 23.6 μg/m³ at AM, 71.8 μg/m³ at BS, and 20.4 μg/m³ at FKL in the entire period. Correlations were moderate to high (0.56–0.64 across sites and datasets). On dust days, MEDIAN continued to give the lowest RMSE while avoiding the large positive biases of NASA-GEOS and NOA-WRF. Figures 9–11 show MEDIAN distributed predictions more evenly for both $O_{PM_{10}}$ and O_{coarse}.

Overall, $O_{PM_{10}}$ was consistently reproduced with higher correlations and tighter clustering in Figures 3–8 compared with Ocoarse, which showed lower correlations and greater scatter. NASA-GEOS and MEDIAN aligned most closely with the observed datasets across periods, NOA-WRF and DREAMABOL added strength at FKL, while EMA-REG4, LOTUS, and SILAM remained the weakest. LOTUS was further limited by its restricted data availability, particularly at FKL.

.The performance of dust forecasting models is critically influenced by several key configuration parameters, as outlined in Table 1 (Model Input Configuration) and

Table 2 (Model Output Configuration). These parameters include vertical resolution, the height of the first layer, data assimilation techniques, horizontal resolution, meteorological drivers, and emission schemes. Each of these factors contributes to the models' ability to accurately predict dust events, which is essential for effective environmental and health risk management.

Models with higher vertical resolution, such as NASA-GEOS and MACC-ECMWF, consistently perform well across various study sites, largely due to their capability to accurately represent the vertical distribution of dust. This feature is especially crucial during long-range transport events, where models need to capture dust movement at different atmospheric levels. For instance, NASA-GEOS achieved strong correlation coefficients at sites like Ayia Marina and Finokalia, underscoring the importance of vertical resolution in complex dust scenarios. In contrast, models with fewer vertical layers, like LOTUS, show limitations in areas with complex topography or high variability in dust inflow. This variability highlights how model design affects performance depending on site characteristics.

Another key factor is the proximity of the first model layer to the ground, which has a direct impact on the accuracy of near-surface dust concentration predictions. Models such as MACC-ECMWF and NCEP-NGAC, with their first layers positioned closer to the surface, show higher correlations with observed particulate matter at surface level. This finding suggests that a lower starting height enhances a model's ability to capture near-surface dust levels accurately—particularly valuable during intense dust events when ground-level concentrations are of greatest concern.

Beyond structural design, data assimilation significantly enhances model performance, especially in adjusting predictions for real-time accuracy. Models incorporating data assimilation, like NASA-GEOS, achieve not only higher correlation coefficients but also reduced biases and RMSEs. This is especially beneficial during extreme dust events, where rapid updates based on real-time data

~~are critical for maintaining predictive accuracy. The effectiveness of data assimilation across models further demonstrates its value in capturing the dynamic nature of dust events.~~

~~Horizontal resolution also plays an important role, as finer grids allow models to improve spatial accuracy, particularly in regions with complex terrain. For instance, models with finer grids, such as NOA-WRF, demonstrate improved ability to handle spatial variability, yielding higher correlation and lower bias at specific sites like Be'er Sheva and Finokalia. This reinforces the need for spatial detail when modelling dust in diverse geographical settings, where elevation and landscape variation influence dust distribution patterns.~~

~~Finally, the choice of meteorological driver and the specific emission schemes employed by models also significantly impact their performance. For example, models driven by ECMWF and GEOS-5 generally outperformed those with other meteorological drivers, demonstrating both higher correlation coefficients and lower biases. Similarly, emission schemes that effectively simulate dust lifting and deposition processes, such as those used in MACC-ECMWF and NASA-GEOS, tend to enhance model performance, particularly in accurately forecasting the intensity and spatial distribution of dust during high-dust events.~~

~~In conclusion, the configuration of dust models plays a crucial role in their predictive accuracy, with vertical resolution, the height of the first layer, data assimilation, horizontal resolution, meteorological drivers, and emission schemes all influencing their effectiveness. Understanding these relationships is crucial for improving model forecasts and effectively managing the environmental and health impacts of dust events in the Eastern Mediterranean region.~~

6. Conclusions and Future Directions in Dust Modelling in EMR

This study provides a comprehensive evaluation of eleven operational dust models and a multi-model ~~ensemble~~ median (~~MMM~~MEDIAN) in forecasting PM concentrations across the EMR. Evaluating operational dust forecasting numerical models against ground measurements is essential to identifying for determining which models perform better in different areas. Such evaluations are crucial for enhancing the accuracy and reliability of DDS - EWS, as these systems rely on robust forecasting models to provide timely alerts for hazardous dust events. Despite their importance, few studies have conducted such evaluations against ground-level measurements, highlighting a critical gap that this research aims to address.

6.1. Key findings

The results clearly indicate that certain models outperform others in predicting ground-level PM concentrations across specific regions, although their performance varies based on the evaluation method applied. Notably, no single model consistently achieves accurate predictions across all three regions, underscoring the need for model-specific adaptations or improvements tailored to regional characteristics. The key findings of this study are summarised as follows:

a) Model Performance Across Sites: The accuracy of individual models varied significantly across the three sites (Ayia Marina, Finokalia, and Be'er Sheva). NASA-GEOS emerged as the most consistent performer, particularly at Ayia Marina and Be'er Sheva, benefiting from a high-resolution vertical structure, data assimilation, and robust meteorological inputs via the GEOS-5 driver. Conversely, models like EMA-REG4 showed substantial limitations, especially in accurately predicting PM levels in Cyprus, indicating that regional drivers and lower-resolution configurations may not be suitable for cross-boundary events in the EMR.

b) Enhancing Early Warning Systems: The findings of this study can provide valuable insights for improving EWS and mitigation strategies in the EMR. By identifying the strengths and weaknesses of the evaluated dust models, particularly in predicting extreme events and PM concentrations, these results can help tailor EWS to the region's specific needs. For example, models such as NASA-GEOS and NOA-WRF, which demonstrated higher accuracy and correlation during intense dust events,

can be prioritized for operational forecasting to provide timely warnings to vulnerable populations. Additionally, the multi-model ~~ensemble~~ median (~~MMM~~MEDIAN), with its reduced uncertainty, can enhance the reliability of forecasts, supporting decision-makers in public health protection and emergency preparedness. These improved predictions are essential for issuing health advisories, minimizing exposure to hazardous PM levels, and informing policy measures to mitigate the adverse impacts of DDS events on air quality and public health.

~~b)~~c) **Differences between 95th percentile dataset and dust days by Achilleos et al.2019:** The results from these 2 datasets especially for the best performing models of NASA GEOS and NOA WRF or the least performing such as EMA REG, are very similar as both datasets are set to describe the highest dust events. Even though they seem similar after the best and least performing models, the rankings vary as the other models are not as consistent as the ones mentioned before. For example, MACC ECMWF and NCEP models have better ranking when the confirmed dust days is used indicating that the multi parameter criteria approach in identifying dust days helps models to improve in their forecasts. Also, the dust days approach lowers the range between values such as RMSE and MB as shown in Tables 4 and 5 and in the BOOT methodology graphs in Figures 6 and 8 for both approaches respectively, thus decreasing the uncertainty between approaches.

~~c)~~d) **Impact of Model Configurations on Performance:**

i. **Meteorological Drivers and Data Assimilation**: ECMWF and GEOS-5-driven models demonstrated higher correlations with observed data, particularly when data assimilation was included. Data assimilation improves models' responsiveness to rapid dust influx and significantly enhances reliability in high-dust scenarios, underscoring its necessity in complex regions like the EMR. Optimizing dust forecasting models is crucial for improving predictive accuracy, especially in regions affected by transboundary dust transport. These advancements align with the goals of initiatives like the Horizon Europe CiROCCO project, which aims to strengthen dust storm monitoring through integrated predictive frameworks.

ii. **Horizontal and Vertical Resolution:** Models with finer horizontal and vertical resolutions, such as NOA-WRF (0.19º x 0.22º, 30 layers) and NASA-GEOS (0.25º x 0.3125º, 72 layers), generally showed improved spatial and temporal accuracy across sites. Enhanced vertical resolution, particularly for capturing dust plumes at altitude, contributed to better event tracking, while horizontal detail supported finer spatial representation critical for sites like Be'er Sheva with localized dust sources.

iii. **First Layer Height and Emission Schemes:** Lower first-layer heights (10-20 m) improved the models' capacity to capture ground-level concentrations accurately, aligning predictions with near-surface PM10 observations. Emission schemes, especially those incorporating sandblasting (e.g., Marticorena and Bergametti, 1995), supported accurate predictions during intense dust events by modelling larger particle transport effectively.

~~d)~~e) **Role of Multi-Model Ensemble (~~MMM~~MEDIAN):** The multi-model ~~ensemble~~median, which averaged predictions across all models, consistently reduced biases and RMSE values, supporting its effectiveness as a robust forecasting operational approach. Although it does not consistently outperform individual models such as NASA-GEOS, the MEDIAN~~MMM~~'s reduced error scores make it a reliable option for broader operational applications in the Eastern Mediterranean region (EMR).

~~e)~~f) **Approaches and methods of evaluation:** The different approaches $O_{PM_{10}}$ and O_{coarse}, different datasets and methods of evaluation have shown that some model's performance can be perceived differently if not properly evaluated by different methods. An example is EMA-REG4, that while being evaluated by performance metrics and BOOT graph methodology it has a very poor performance but on the contingency tables it has one of the best overall scores. Furthermore, swapping from $O_{PM_{10}}$ to O_{coarse}, it shows how sensitive are models to fine materials as they do or do not take them into account changing an underestimation to overestimation and vice verca, while their correlation is mostly decreasing.

This study has limitations, including the absence of recent $PM_{2.5}$ data in some regions and potential changes in model configurations that might affect performance. Despite these limitations, operational models remain reliable for predicting DDS events and provide

valuable mitigation insights (Eleftheriou et al., 2023). However, continuous technological advancements are essential to enhance dust prediction capabilities, particularly concerning the duration and intensity of DDS events.

6.2. Future Directions

To advance dust forecasting in the Eastern Mediterranean Region (EMR), targeted model improvements and data integration strategies are essential. Enhancing data assimilation through high-frequency satellite data, such as MODIS AOD, and incorporating localized PM monitoring can improve real-time prediction accuracy, particularly when dynamic field measurements like $PM_{2.5}$ data are included. Importantly, expanding data assimilation to include more in-situ measurements by strengthening monitoring networks, especially in underrepresented source areas, would enhance the accuracy of model predictions and ensure better representation of dust transport from these critical regions. Also, sometimes the models fail producing negative or even zero concentration thus influencing the evaluation technique used to assess their performance. Such results should be redefined to a Limit of Detection (LOD) consistent with observation techniques to provide a more realistic result and the result to be valid in order to be evaluated.

The variability in prediction accuracy across models also highlights the need for improved, site-specific emission parameterizations. Adaptive emission schemes that account for local soil characteristics, vegetation cover, and land use would be especially beneficial for urban-adjacent and desert-proximal sites like Be'er Sheva. Furthermore, results consistently favouring models with finer spatial (0.1°–0.3°) and vertical resolutions (over 50 layers) suggest that these refinements are crucial for accurately capturing dust dynamics and transboundary events impacting EMR air quality.

Customizing multi-model ~~ensemble~~ median strategies by including only the top-performing models for each site could enhance forecast precision while preserving the bias reduction advantages seen in multi-model ~~ensembles~~medians. Such an approach leverages specific model strengths under varied conditions, especially during peak dust events. Expanding model validation efforts to include diverse datasets—such as vertical aerosol profiles, aerosol chemistry, and source-specific particle data—would further refine model parameterizations and improve reliability. Collaborating with regional networks to utilize observational data from North Africa, the Middle East, and EMR could substantially strengthen model robustness.

Future efforts should focus on refining model parameterizations, integrating localized field data, and validating models against diverse datasets to ensure predictive robustness and reliability. Overall, the results underscore the region-specific nature of dust forecasting and highlight model configurations and features that enhance accuracy at individual sites.

Code and Data Availability.

Model forecast data from the SDS-WAS Barcelona Dust Forecast Center (https://sds-was.aemet.es/) are openly accessible and were archived during the analysis period. The ground-based measurement data used in this study are available upon request from the authors. This study was performed using custom scripts in MATLAB (more details can be found at https://www.mathworks.com) and R (it is available at https://www.r-project.org). While no new modelling code was developed in this study, the scripts used for data processing, correlation analysis, and performance evaluation are available upon request from the corresponding author.

Author contributions.

MNKA, PM, and AE contributed to manuscript preparation. MNKA and PY provided the financial support for this study. MNKA also contributed to the investigation, provided resources, and participated in the supervision of the study. PM had the conceptualization, investigation, formal analysis, and supervision of this study. AE and PM performed the formal analysis of the

data; AE also prepared the original draft of the manuscript. NK, IK, EV, and CS were responsible for data curation and contributed to manuscript review and editing. PK and PY contributed to writing – review and editing. All co-authors reviewed and approved the final version of the manuscript.

Competing interests.

The contact author has declared that none of the authors has any competing interests.

Acknowledgments

This research work is part of the CiROCCO Project, funded by the European Union (Grant no. 101086497). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or REA. Neither the European Union nor the granting authority can be held responsible for them. Moreover, the authors are grateful for the financial support from the LIFE+-MEDEA Program under Grant Agreement LIFE16 CCA/CY/000041. The authors would also like to express their deepest gratitude to the data providers, the organizations, and their researchers who develop and operate the operational models, as well as to the SDS-WAS portal for the collection and availability of the data.

References

Achilleos, S., Al-Ozairi, E., Alahmad, B., Garshick, E., Neophytou, A. M., Bouhamra, W., Yassin, M. F., and Koutrakis, P.: Acute effects of air pollution on mortality: A 17-year analysis in Kuwait, Environment International, 126, 476–483, https://doi.org/10.1016/j.envint.2019.01.072, 2019.

Achilleos, S., Mouzourides, P., Kalivitis, N., Katra, I., Kloog, I., Kouis, P., Middleton, N., Mihalopoulos, N., Neophytou, M., Panayiotou, A., Papatheodorou, S., Savvides, C., Tymvios, F., Vasiliadou, E., Yiallouros, P., and Koutrakis, P.: Spatio-temporal variability of desert dust storms in Eastern Mediterranean (Crete, Cyprus, Israel) between 2006 and 2017 using a uniform methodology, Science of The Total Environment, 714, 136693, https://doi.org/10.1016/j.scitotenv.2020.136693, 2020.

Achilleos, S., Michanikou, A., Kouis, P., Papatheodorou, S. I., Panayiotou, A. G., Kinni, P., Mihalopoulos, N., Kalivitis, N., Kouvarakis, G., Galanakis, E., Michailidi, E., Tymvios, F., Chrysanthou, A., Neophytou, M., Mouzourides, P., Savvides, C., Vasiliadou, E., Papasavvas, I., Christophides, T., Nicolaou, R., Avraamides, P., Kang, C.-M., Middleton, N., Koutrakis, P., and Yiallouros, P. K.: Improved indoor air quality during desert dust storms: The impact of the MEDEA exposure-reduction strategies, Science of The Total Environment, 863, 160973, https://doi.org/10.1016/j.scitotenv.2022.160973, 2023.

Alfaro, S. C. and Gomes, L.: Modeling mineral aerosol production by wind erosion: Emission intensities and aerosol size distributions in source areas, J. Geophys. Res., 106, 18075–18084, https://doi.org/10.1029/2000JD900339, 2001.

Basart, S., Pérez, C., Nickovic, S., Cuevas, E., and Baldasano, J. M.: Development and evaluation of the BSC-DREAM8b dust regional model over Northern Africa, the Mediterranean and the Middle East, Tellus B: Chemical and Physical Meteorology, 64, 18539, https://doi.org/10.3402/tellusb.v64i0.18539, 2012.

Bodenheimer, S., Lensky, I. M., and Dayan, U.: Characterization of Eastern Mediterranean dust storms by area of origin; North Africa vs. Arabian Peninsula, Atmospheric Environment, 198, 158–165, https://doi.org/10.1016/j.atmosenv.2018.10.034, 2019.

Çapraz, Ö. and Deniz, A.: Particulate matter (PM10 and PM2.5) concentrations during a Saharan dust episode in Istanbul, Air Qual Atmos Health, 14, 109–116, https://doi.org/10.1007/s11869-020-00917-4, 2021.

Chang, J. C. and Hanna, S. R.: Air quality model performance evaluation, Meteorol Atmos Phys, 87, https://doi.org/10.1007/s00703-003-0070-7, 2004.

~~Chen, J. and Hoek, G.: Long-term exposure to PM and all-cause and cause-specific mortality: A systematic review and meta-analysis, Environment International, 143, 105974, https://doi.org/10.1016/j.envint.2020.105974, 2020.~~

Colarco, P., Da Silva, A., Chin, M., and Diehl, T.: Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth, J. Geophys. Res., 115, 2009JD012820, https://doi.org/10.1029/2009JD012820, 2010.

Eleftheriou, A., Mouzourides, P., Biskos, G., Yiallouros, P., Kumar, P., and Neophytou, M. K.-A.: The challenge of adopting mitigation and adaptation measures for the impacts of sand and dust storms in Eastern Mediterranean Region: a critical review, Mitig Adapt Strateg Glob Change, 28, 33, https://doi.org/10.1007/s11027-023-10070-9, 2023.

Flaounas, E., Kotroni, V., Lagouvardos, K., Klose, M., Flamant, C., and Giannaros, T. M.: Sensitivity of the WRF-Chem (V3.6.1) model to different dust emission parametrisation: assessment in the broader Mediterranean region, Geosci. Model Dev., 10, 2925–2945, https://doi.org/10.5194/gmd-10-2925-2017, 2017.

García-Castrillo, G. and E. Terradellas: Evaluation of the dust Forecasts in the Canary Islands, WMO SDS-WAS, Barcelona, 21 pp. SDS-WAS-2017-002, 2017.

Ginoux, P., Chin, M., Tegen, I., Prospero, J. M., Holben, B., Dubovik, O., and Lin, S.: Sources and distributions of dust aerosols simulated with the GOCART model, J. Geophys. Res., 106, 20255–20273, https://doi.org/10.1029/2000JD000053, 2001.

Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y. J., Nakajima, T., Lavenu, F., Jankowiak, I., and Smirnov, A.: AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization, Remote Sensing of Environment, 66, 1–16, https://doi.org/10.1016/S0034-4257(98)00031-5, 1998.

Janjić, Z. I.: The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes, Mon. Wea. Rev., 122, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2, 1994.

Jones, B. A.: Dust storms and human well-being, Resource and Energy Economics, 72, 101362, https://doi.org/10.1016/j.reseneeco.2023.101362, 2023.

Kazadzis, S., Bais, A., Balis, D., Kouremeti, N., Zempila, M., Arola, A., Giannakaki, E., Amiridis, V., and Kazantzidis, A.: Spatial and temporal UV irradiance and aerosol variability within the area of an OMI satellite pixel, Atmos. Chem. Phys., 9, 4593–4601, https://doi.org/10.5194/acp-9-4593-2009, 2009.

Klose, M., Jorba, O., Gonçalves Ageitos, M., Escribano, J., Dawson, M. L., Obiso, V., Di Tomaso, E., Basart, S., Montané Pinto, G., Macchia, F., Ginoux, P., Guerschman, J., Prigent, C., Huang, Y., Kok, J. F., Miller, R. L., and Pérez García-Pando, C.: Mineral dust cycle in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (MONARCH) Version 2.0, Geosci. Model Dev., 14, 6403–6444, https://doi.org/10.5194/gmd-14-6403-2021, 2021.

Knippertz, P. and Stuut, J.-B. W. (Eds.): Mineral Dust: A Key Player in the Earth System, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-017-8978-3, 2014.

Krasnov, H., Katra, I., and Friger, M.: Increase in dust storm related PM10 concentrations: A time series analysis of 2001–2015, Environmental Pollution, 213, 36–42, https://doi.org/10.1016/j.envpol.2015.10.021, 2016.

Lahiri, S. N. (Ed.): Resampling methods for dependent data, Springer, New York Berlin Heidelberg, 374 pp., 2010.

Lorentzou, C., Kouvarakis, G., Kozyrakis, G. V., Kampanis, N. A., Trahanatzi, I., Fraidakis, O., Tzanakis, N., Kanakidou, M., Agouridakis, P., and Notas, G.: Extreme desert dust storms and COPD morbidity on the island of Crete, COPD, Volume 14, 1763–1768, https://doi.org/10.2147/COPD.S208108, 2019.

Lu, C.-H., Da Silva, A., Wang, J., Moorthi, S., Chin, M., Colarco, P., Tang, Y., Bhattacharjee, P. S., Chen, S.-P., Chuang, H.-Y., Juang, H.-M. H., McQueen, J., and Iredell, M.: The implementation of NEMS GFS Aerosol Component (NGAC) Version 1.0 for global dust forecasting at NOAA/NCEP, Geosci. Model Dev., 9, 1905–1919, https://doi.org/10.5194/gmd-9-1905-2016, 2016.

Lumet, E., Jaravel, T., Rochoux, M. C., Vermorel, O., and Lacroix, S.: Assessing the Internal Variability of Large-Eddy Simulations for Microscale Pollutant Dispersion Prediction in an Idealized Urban Environment, Boundary-Layer Meteorol, 190, 9, https://doi.org/10.1007/s10546-023-00853-7, 2024.

Lwin, K. S., Tobias, A., Chua, P. L., Yuan, L., Thawonmas, R., Ith, S., Htay, Z. W., Yu, L. S., Yamasaki, L., Roqué, M., Querol, X., Fussell, J. C., Nadeau, K. C., Stafoggia, M., Saliba, N. A., Sheng Ng, C. F., and Hashizume, M.: Effects of Desert Dust and Sandstorms on Human Health: A Scoping Review, GeoHealth, 7, e2022GH000728, https://doi.org/10.1029/2022GH000728, 2023.

Manders, A. M. M., Builtjes, P. J. H., Curier, L., Denier Van Der Gon, H. A. C., Hendriks, C., Jonkers, S., Kranenburg, R., Kuenen, J. J. P., Segers, A. J., Timmermans, R. M. A., Visschedijk, A. J. H., Wichink Kruit, R. J., Van Pul, W. A. J., Sauter, F. J., Van Der Swaluw, E., Swart, D. P. J., Douros, J., Eskes, H., Van Meijgaard, E., Van Ulft, B., Van Velthoven, P., Banzhaf, S., Mues, A. C., Stern, R., Fu, G., Lu, S., Heemink, A., Van Velzen, N., and Schaap, M.: Curriculum vitae of the LOTOS–EUROS (v2.0) chemistry transport model, Geosci. Model Dev., 10, 4145–4173, https://doi.org/10.5194/gmd-10-4145-2017, 2017.

Marticorena, B. and Bergametti, G.: Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme, J. Geophys. Res., 100, 16415–16430, https://doi.org/10.1029/95JD00690, 1995.

Mircea, M., D'Isidoro, M., Maurizi, A., Vitali, L., Monforti, F., Zanini, G., and Tampieri, F.: A comprehensive performance evaluation of the air quality model BOLCHEM to reproduce the ozone concentrations over Italy, Atmospheric Environment, 42, 1169–1185, https://doi.org/10.1016/j.atmosenv.2007.10.043, 2008.

Morcrette, J. -J., Beljaars, A., Benedetti, A., Jones, L., and Boucher, O.: Sea-salt and dust aerosols in the ECMWF IFS model, Geophysical Research Letters, 35, 2008GL036041, https://doi.org/10.1029/2008GL036041, 2008.

Morcrette, J. -J., Boucher, O., Jones, L., Salmond, D., Bechtold, P., Beljaars, A., Benedetti, A., Bonet, A., Kaiser, J. W., Razinger, M., Schulz, M., Serrar, S., Simmons, A. J., Sofiev, M., Suttie, M., Tompkins, A. M., and Untch, A.: Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: Forward modeling, J. Geophys. Res., 114, 2008JD011235, https://doi.org/10.1029/2008JD011235, 2009.

Mouzourides, P., Kumar, P., and Neophytou, M. K.-A.: Assessment of long-term measurements of particulate matter and gaseous pollutants in South-East Mediterranean, Atmospheric Environment, 107, 148–165, https://doi.org/10.1016/j.atmosenv.2015.02.031, 2015.

NASA Publications: MODIS, https://terra.nasa.gov/about/terra-instruments/modis (last access 27. February, 2025), 2025.

Nickovic, S., Kallos, G., Papadopoulos, A., and Kakaliagou, O.: A model for prediction of desert dust cycle in the atmosphere, J. Geophys. Res., 106, 18113–18129, https://doi.org/10.1029/2000JD900794, 2001.

Nickovic, S., Cvetkovic, B., Madonna, F., Rosoldi, M., Pejanovic, G., Petkovic, S., and Nikolic, J.: Cloud ice caused by atmospheric mineral dust – Part 1: Parameterization ofice nuclei concentration in the NMME-DREAM model, Atmos. Chem. Phys., 16, 11367–11378, https://doi.org/10.5194/acp-16-11367-2016, 2016.

Pérez, C., Nickovic, S., Pejanovic, G., Baldasano, J. M., and Özsoy, E.: Interactive dust-radiation modeling: A step to improve weather forecasts, J. Geophys. Res., 111, 2005JD006717, https://doi.org/10.1029/2005JD006717, 2006.

Pérez, C., Haustein, K., Janjic, Z., Jorba, O., Huneeus, N., Baldasano, J. M., Black, T., Basart, S., Nickovic, S., Miller, R. L., Perlwitz, J. P., Schulz, M., and Thomson, M.: Atmospheric dust modeling from meso to global scales with the online NMMB/BSC-Dust model – Part 1: Model description, annual simulations and evaluation, https://doi.org/10.5194/acpd-11-17551-2011, 22 June 2011.

Querol, X., Pey, J., Pandolfi, M., Alastuey, A., Cusack, M., Pérez, N., Moreno, T., Viana, M., Mihalopoulos, N., Kallos, G., and Kleanthous, S.: African dust contributions to mean ambient PM10 mass-levels across the Mediterranean Basin, Atmospheric Environment, 43, 4266–4277, https://doi.org/10.1016/j.atmosenv.2009.06.013, 2009.

Rémy, S., Kipling, Z., Huijnen, V., Flemming, J., Nabat, P., Michou, M., Ades, M., Engelen, R., and Peuch, V.-H.: Description and evaluation of the tropospheric aerosol scheme in the Integrated Forecasting System (IFS-AER, cycle 47R1) of ECMWF, Geosci. Model Dev., 15, 4881–4912, https://doi.org/10.5194/gmd-15-4881-2022, 2022.

Shao, Y., Raupach, M. R., and Findlater, P. A.: Effect of saltation bombardment on the entrainment of dust by wind, J. Geophys. Res., 98, 12719–12726, https://doi.org/10.1029/93JD00396, 1993.

Sofiev, M., Vira, J., Kouznetsov, R., Prank, M., Soares, J., and Genikhovich, E.: Construction of the SILAM Eulerian atmospheric dispersion model based on the advection algorithm of Michael Galperin, Geosci. Model Dev., 8, 3497–3522, https://doi.org/10.5194/gmd-8-3497-2015, 2015.

Solomos, S., Spyrou, C., Barreto, A., Rodríguez, S., González, Y., Neophytou, M. K. A., Mouzourides, P., Bartsotas, N. S., Kalogeri, C., Nickovic, S., Vukovic Vimic, A., Vujadinovic Mandic, M., Pejanovic, G., Cvetkovic, B., Amiridis, V., Sykioti, O., Gkikas, A., and Zerefos, C.: The Development of METAL-WRF Regional Model for the Description of Dust Mineralogy in the Atmosphere, Atmosphere, 14, 1615, https://doi.org/10.3390/atmos14111615, 2023.

United Nations Publications: https://press.un.org/en/2024/ga12613.doc.htm#:~:text=The%20193%2Dmember%20organ%20first,fight%20against%20those%20meteorological%20phenomena (last access 27. February, 2025), 2024.

Tegen, I. and Fung, I.: Modeling of mineral dust in the atmosphere: Sources, transport, and optical thickness, J. Geophys. Res., 99, 22897–22914, https://doi.org/10.1029/94JD01928, 1994.

Triantafyllou, E., Diapouli, E., Korras-Carraca, M. B., Manousakas, M., Psanis, C., Floutsi, A. A., Spyrou, C., Eleftheriadis, K., and Biskos, G.: Contribution of locally-produced and transported air pollution to particulate matter in a small insular coastal city, Atmospheric Pollution Research, 11, 667–678, https://doi.org/10.1016/j.apr.2019.12.015, 2020.

Tsiflikiotou, M. A., Kostenidou, E., Papanastasiou, D. K., Patoulias, D., Zarmpas, P., Paraskevopoulou, D., Diapouli, E., Kaltsonoudis, C., Florou, K., Bougiatioti, A., Stavroulas, I., Theodosi, C., Kouvarakis, G., Vasilatou, V., Siakavaras, D., Biskos, G., Pilinis, C., Eleftheriadis, K., Gerasopoulos, E., Mihalopoulos, N., and Pandis, S. N.: Summertime particulate matter and its composition in Greece, Atmospheric Environment, 213, 597–607, https://doi.org/10.1016/j.atmosenv.2019.06.013, 2019.

Vratolis, S., Gini, M. I., Bezantakos, S., Stavroulas, I., Kalivitis, N., Kostenidou, E., Louvaris, E., Siakavaras, D., Biskos, G., Mihalopoulos, N., Pandis, S. N., Pilinis, C., Papayannis, A., and Eleftheriadis, K.: Particle number size distribution statistics at City-Centre Urban Background, urban background, and remote stations in Greece during summer, Atmospheric Environment, 213, 711–726, https://doi.org/10.1016/j.atmosenv.2019.05.064, 2019.

Yang, L., Fang, S., Zhuang, S., Chen, Y., Li, X., and Zhang, Q.: Atmospheric 137Cs dispersion following the Fukushima Daiichi nuclear accident: Local-scale simulations using CALMET and LAPMOD, Annals of Nuclear Energy, 195, 110137, https://doi.org/10.1016/j.anucene.2023.110137, 2024.

Zakey, A. S., Solmon, F., and Giorgi, F.: Implementation and testing of a desert dust module in a regional climate model, Atmos. Chem. Phys., 6, 4687–4704, https://doi.org/10.5194/acp-6-4687-2006, 2006.

Zender, C. S., Bian, H., and Newman, D.: Mineral Dust Entrainment and Deposition (DEAD) model: Description and 1990s dust climatology, J. Geophys. Res., 108, 2002JD002775, https://doi.org/10.1029/2002JD002775, 2003.

Zhang, L., Montuoro, R., McKeen, S. A., Baker, B., Bhattacharjee, P. S., Grell, G. A., Henderson, J., Pan, L., Frost, G. J., McQueen, J., Saylor, R., Li, H., Ahmadov, R., Wang, J., Stajner, I., Kondragunta, S., Zhang, X., and Li, F.: Development and evaluation of the Aerosol Forecast Member in the National Center for Environment Prediction (NCEP)'s Global Ensemble Forecast System (GEFS-Aerosols v1), Geosci. Model Dev., 15, 5337–5369, https://doi.org/10.5194/gmd-15-5337-2022, 2022.