

Poprawa modelowania dobowych stężeń $PM_{2.5}$ z użyciem uczenia maszynowego

Autorzy:

Tetiana Vovk, Maciej Kryza, Małgorzata Werner

Wydział Nauk o Ziemi i Kształtowania Środowiska

Uniwersytet Wrocławski



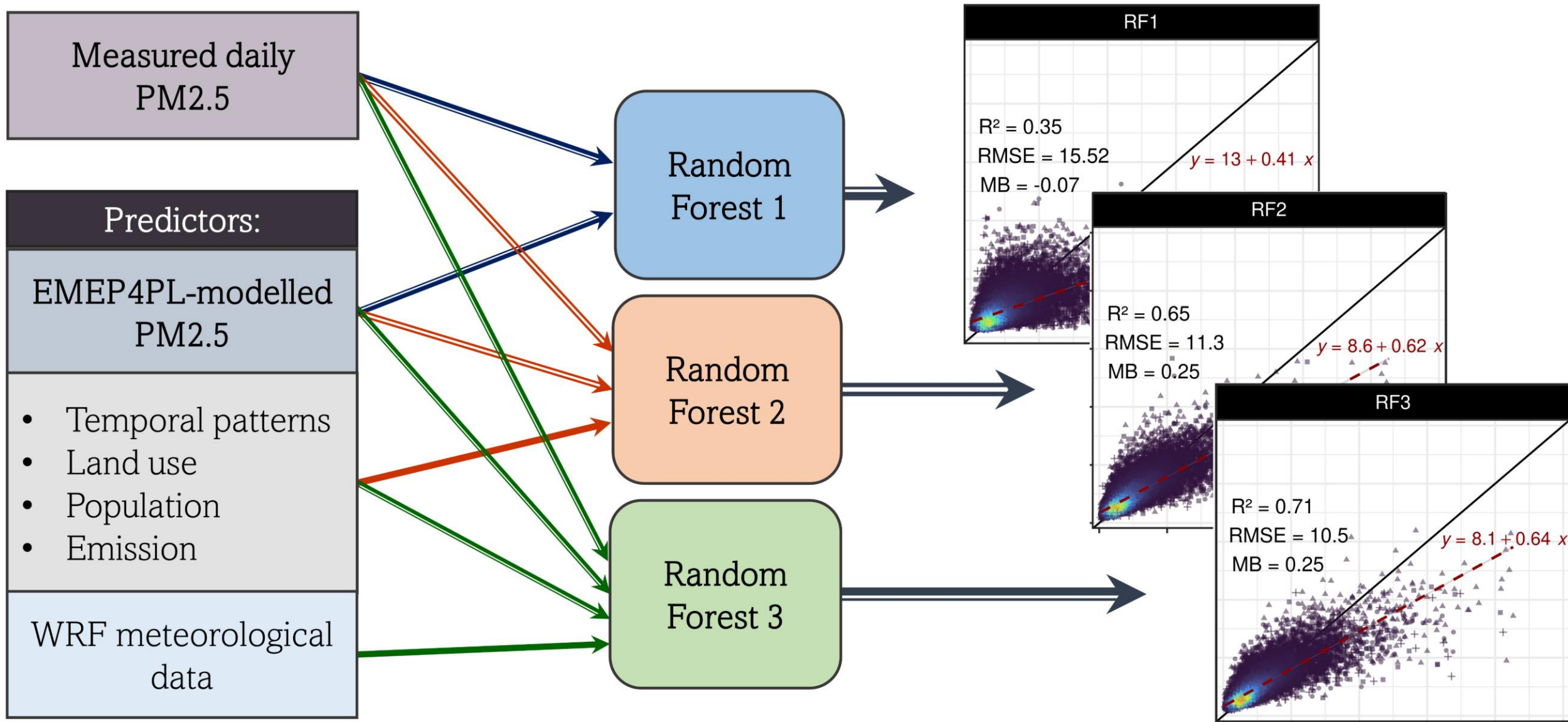
Idea i cel badania

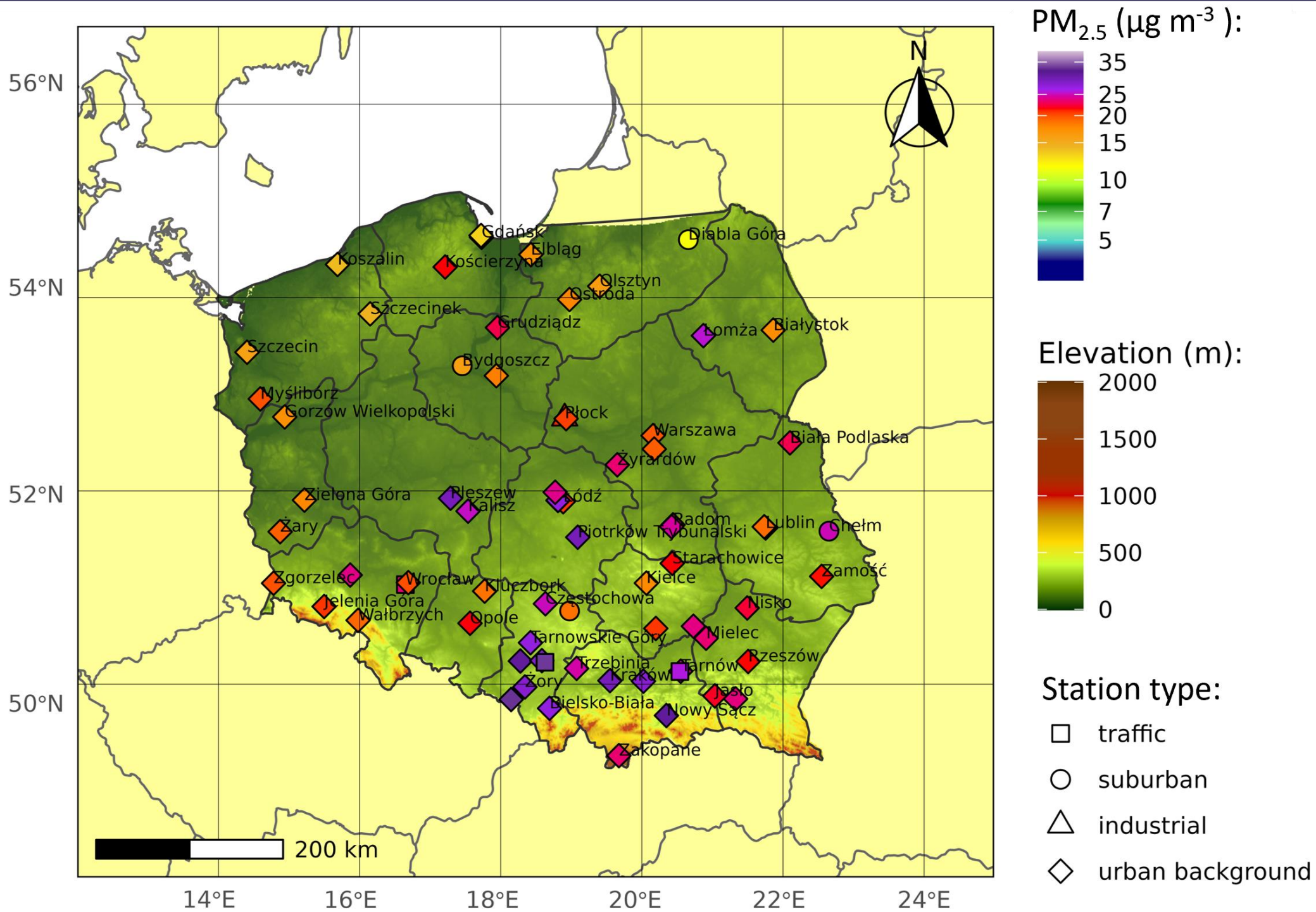
- Wiele istniejących modeli numerycznych w środkowej części Europy wykazuje niższe, niż mierzone na stacjach, poziomy stężenie $PM_{2.5}$
- Lepsza jakość modelowania pozwoli na dokonanie dokładnych szacunków narażenia populacji i wpływu zanieczyszczeń na zdrowie
- Uniwersalna metoda może zostać z sukcesem zaaplikowana w połączeniu z dowolnym modelem numerycznym w regionie

Dane i metody

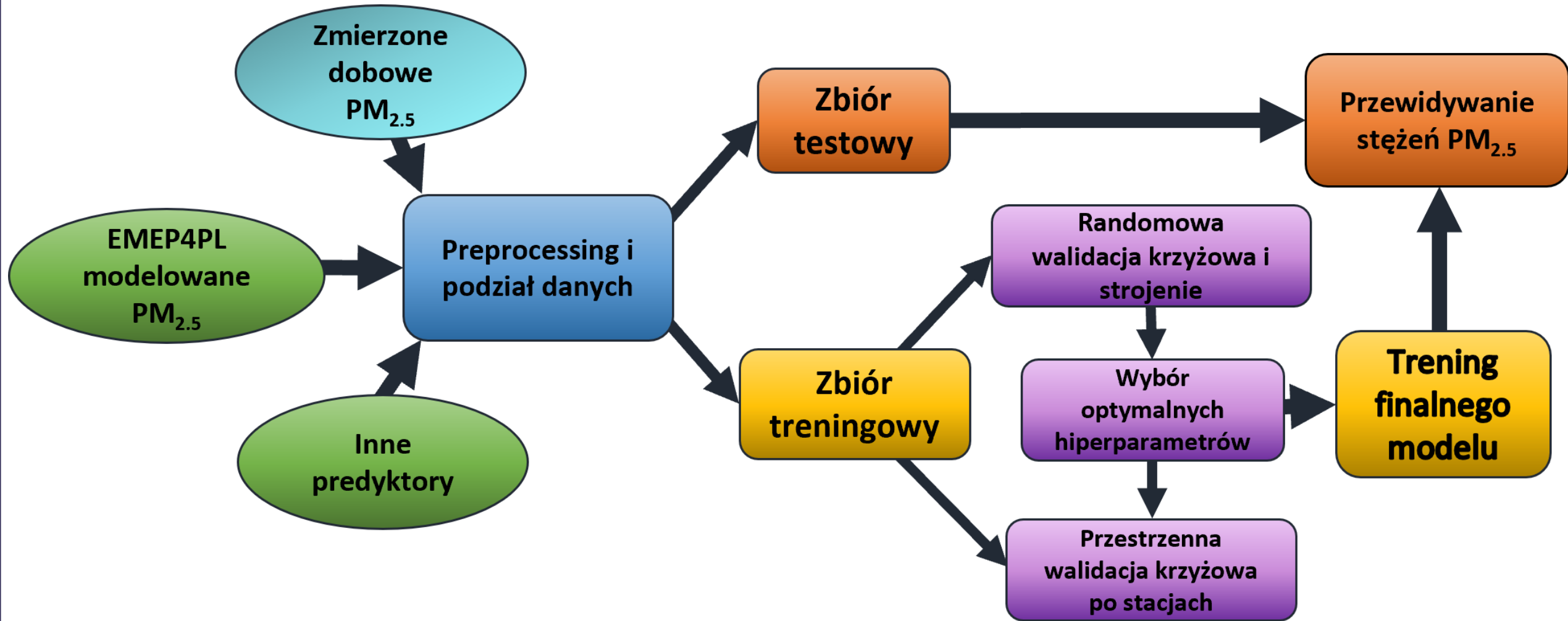
- ✓ dane historyczne $PM_{2.5}$ z 4 lat dla 71 stacji pomiarowych w całej Polsce (GIOŚ)
- ✓ dane przestrzenne stężeń $PM_{2.5}$ z modelu numerycznego EMEP4PL
- ✓ zmienne meteorologiczne z mezoskalowego modelu pogody WRF
- ✓ dane o populacji, emisji i zagospodarowaniu terenu (Eurostat, KOBIZE, Copernicus)
- ✓ przetestowano 3 modele oparte o algorytm używający zespołu drzew decyzyjnych (Random Forest) z różnymi zestawami zmiennych predykcyjnych
- ✓ wyniki opublikowano:

Vovk, T., Kryza, M., & Werner, M. (2024). *Using random forest to improve EMEP4PL model estimates of daily $PM_{2.5}$ in Poland*. *Atmospheric Environment*, 332, 120615.
<https://doi.org/10.1016/j.atmosenv.2024.120615>

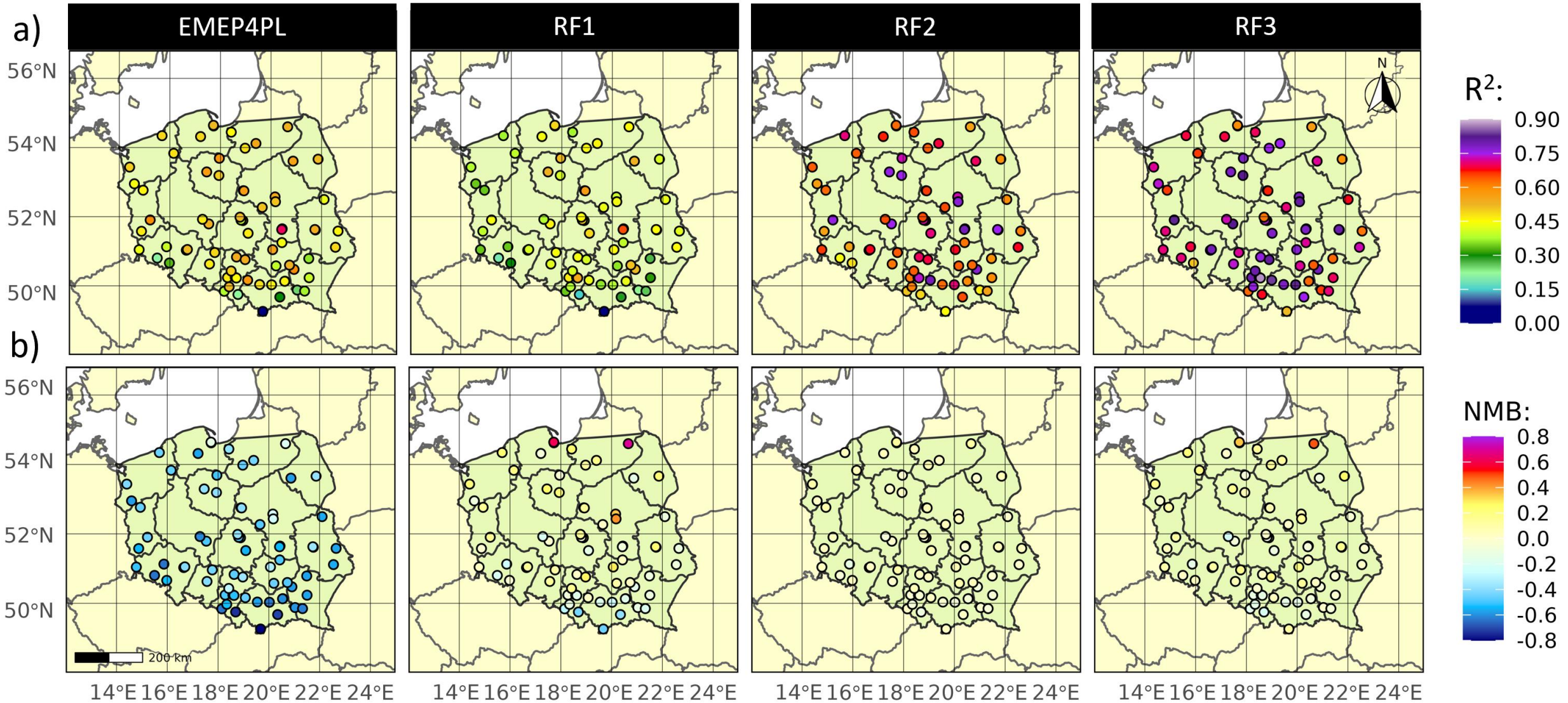




Schemat modelowania

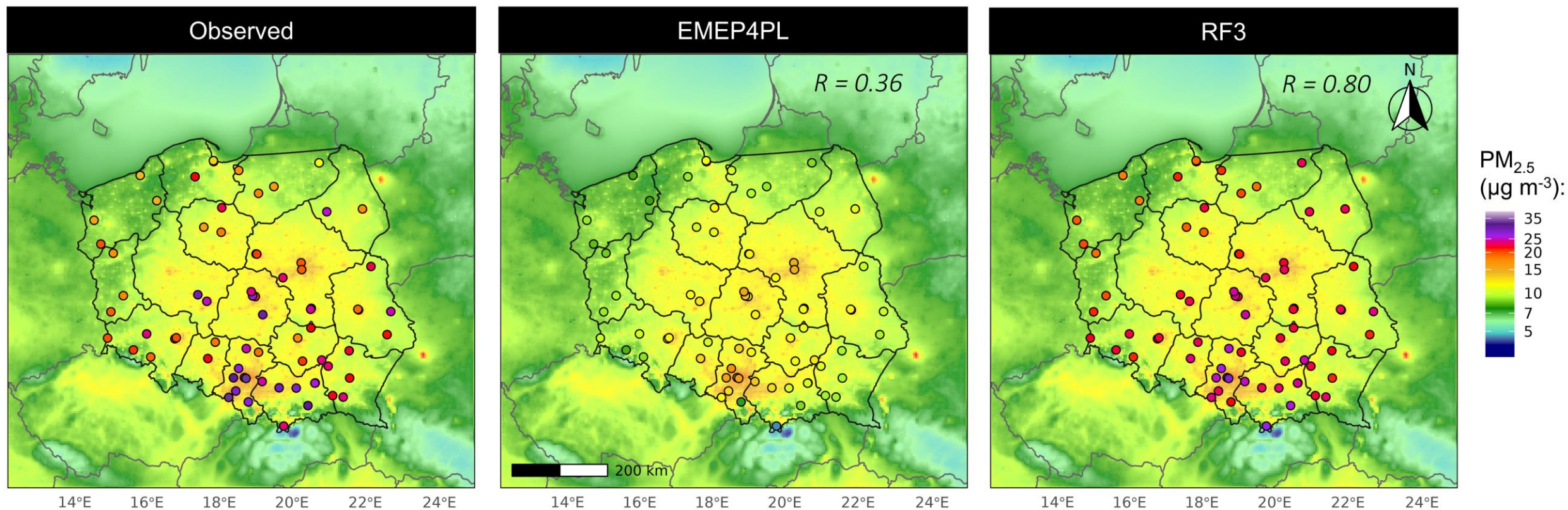


Główne wyniki



Rysunek: Porównanie R^2 (a) i znormalizowanego biasu średniego (NMB) (b) dla EMEP4PL i trzech modeli Random Forest (RF1, RF2, RF3) dla wszystkich stacji (zbiór testowy)

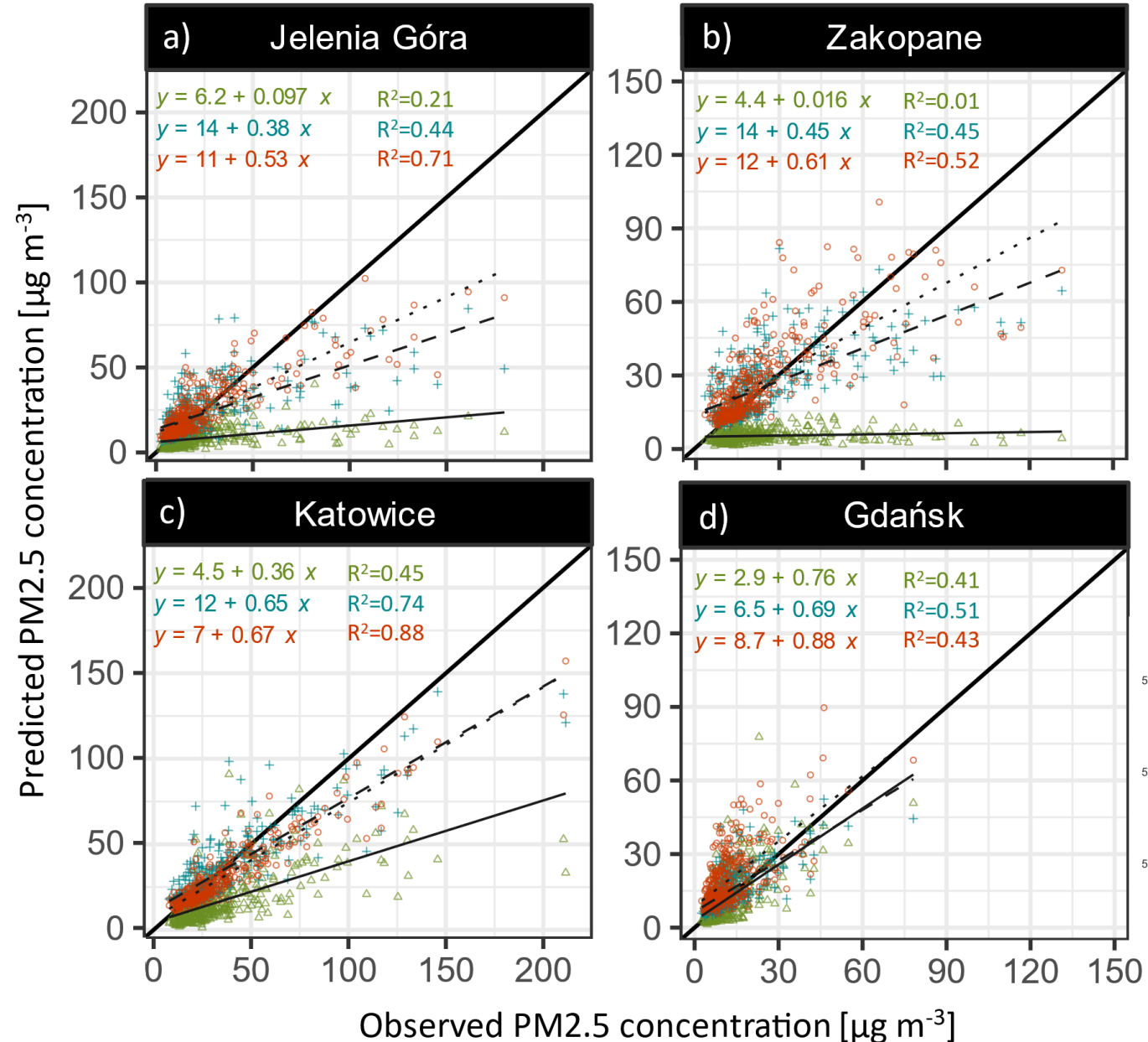
Główne wyniki



Rysunek: Porównanie uśrednionych stężeń PM_{2.5} dla całego zbioru danych na stacjach: obserwowanych, modelowanych przez EMEP4PL i modelowanych przez RF3. Tło rysunku to uśredniona warstwa stężeń PM_{2.5} z EMEP4PL dla całego okresu (2016-2019). Skala kolorów dla warstwy EMEP4PL i dla stacji monitoringowych jest wspólna.

Główne wyniki

Rysunek: Porównanie oszacowań modeli EMEP4PL, RF2 i RF3 na zbiorze testowym dla stacji monitoringowych w Jeleniej Górze (a), Zakopanem (b), Katowicach (c) i Gdańsku (d)



Model:

a Δ EMEP4PL

a + RF2

a \circ RF3

Regression line for model:

— EMEP4PL

- - RF2

... RF3



Podsumowanie

- Uczenie maszynowe skutecznie poprawiło wyniki modelu numerycznego EMEP4PL, również dla obszarów, gdzie EMEP4PL miał najniższą dokładność (stacje podgórskie)
- Niedoszacowanie zostało znacznie obniżone, także dla dni o wysokich stężeniach $PM_{2.5}$ ($>50 \text{ ug m}^{-3}$)
- Random Forest (RF3) z pełnym zestawem zmiennych pokazał największą dokładność ($R^2 = 0.71$, $MB=0.25$), choć RF2, gdzie nie wykorzystano zmiennych meteorologicznych, również znacząco poprawił wyniki modelowania EMEP4PL ($R^2 = 0.65$, $MB = 0.25$)

Dziękuję za uwagę!

Vovk, T., Kryza, M., & Werner, M. (2024). *Using random forest to improve EMEP4PL model estimates of daily PM2.5 in Poland*. *Atmospheric Environment*, 332, 120615. <https://doi.org/10.1016/j.atmosenv.2024.120615>

