

# Supplementary File S1

## EnvChemMetaFlow Complete R Workflow Code

*Uploadable version for supplementary materials*

### Purpose

This supplementary document provides the complete annotated R workflow used to generate the synthetic dataset, perform random-effects meta-analysis, and create the forest plot, funnel plot, PRISMA-style diagram, pollutant distribution heatmap, meta-regression moderator plot, and subgroup forest plot for the EnvChemMetaFlow demonstration manuscript.

### Ethical and Scientific Use Statement

All datasets generated by this code are synthetic and are intended only for methodological demonstration, teaching, workflow validation, and reproducibility testing. The generated estimates, plots, and statistical outputs must not be interpreted as real environmental monitoring results or used as evidence for environmental policy decisions.

### How to Use This File

Copy the R code below into an R script or RStudio session and run it in a working directory where output files should be saved. The script will generate spreadsheet outputs, PNG/PDF figures, and text outputs for meta-analysis and moderator analysis.

### Complete R Code

```
# =====
# Supplementary File S1
# EnvChemMetaFlow: Complete Uploadable R Workflow Code
# =====
# This supplementary computational workflow contains the R code for:
# 1. Synthetic dataset generation
# 2. Random-effects meta-analysis
# 3. Forest plot and funnel plot generation
# 4. Egger's regression test
# 5. Pollutant distribution heatmap
# 6. PRISMA-style workflow diagram
# 7. Meta-regression moderator analysis
# 8. Subgroup forest plot by water type
#
# IMPORTANT ETHICAL NOTE:
# All generated data are synthetic and intended only for workflow
# demonstration, teaching, and reproducibility testing. They must not
# be interpreted as real environmental monitoring evidence.
# =====

# =====
# PART A. CORE SYNTHETIC DATASET AND META-ANALYSIS WORKFLOW
# =====

# =====
# EnvChemMetaFlow Supplementary R Script
# Synthetic Environmental Chemical Data Management and Meta-analysis
# =====
# Purpose:
# This script generates a synthetic environmental chemical monitoring dataset
# and produces publication-style forest and funnel plots for methodological
# demonstration of the EnvChemMetaFlow workflow.
#
# Important note:
# All data generated by this script are synthetic and are intended only for
# workflow demonstration, teaching, and reproducibility testing. They should
# not be interpreted as real environmental monitoring data.
# =====
```

```

# -----
# 1. Load required packages
# -----
required_packages <- c("dplyr", "ggplot2", "readr", "openxlsx", "metafor")

for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
}

library(dplyr)
library(ggplot2)
library(readr)
library(openxlsx)
library(metafor)

# -----
# 2. Set reproducibility seed
# -----
set.seed(42)

# -----
# 3. Generate synthetic dataset
# -----
countries <- c("Bangladesh", "India", "China", "Pakistan", "Vietnam", "Thailand")
compounds <- c("Fluoxetine", "Ciprofloxacin", "Nitrate", "Mixed Pharmaceuticals")
water_types <- c("Groundwater", "Surface Water", "Wastewater")

n_studies <- 32

synthetic_data <- data.frame(
  StudyID = paste0("S", seq_len(n_studies)),
  Country = sample(countries, n_studies, replace = TRUE),
  Compound = sample(compounds, n_studies, replace = TRUE),
  WaterType = sample(water_types, n_studies, replace = TRUE)
)

synthetic_data <- synthetic_data %>%
  rowwise() %>%
  mutate(
    Mean_ng_L = case_when(
      Compound == "Fluoxetine" ~ max(rnorm(1, mean = 50, sd = 15), 5),
      Compound == "Ciprofloxacin" ~ max(rnorm(1, mean = 130, sd = 35), 5),
      Compound == "Nitrate" ~ max(rnorm(1, mean = 4200, sd = 900), 5),
      Compound == "Mixed Pharmaceuticals" ~ max(rnorm(1, mean = 180, sd = 50), 5),
      TRUE ~ NA_real_
    ),
    SD = Mean_ng_L * runif(1, min = 0.12, max = 0.25),
    N = sample(15:80, 1)
  ) %>%
  ungroup() %>%
  mutate(
    Mean_ng_L = round(Mean_ng_L, 2),
    SD = round(SD, 2)
  )

# Save synthetic dataset
write.xlsx(
  synthetic_data,
  file = "EnvChemMetaFlow_Synthetic_Dataset.xlsx",
  overwrite = TRUE
)

# -----
# 4. Prepare data for meta-analysis
# -----
# Because the synthetic dataset includes pollutants on very different scales
# such as pharmaceutical residues and nitrate, log-transformed mean concentration
# is used as the demonstration effect size.

meta_data <- synthetic_data %>%
  mutate(
    Study_Label = paste(StudyID, Country, Compound, sep = " | "),
    yi = log(Mean_ng_L),
    sei = SD / (Mean_ng_L * sqrt(N)),
    vi = sei^2
  )

```

```

# Random-effects model using REML
res <- rma(
  yi = yi,
  sei = sei,
  data = meta_data,
  method = "REML"
)

summary(res)

# Add confidence intervals and weights
meta_data <- meta_data %>%
  mutate(
    ci_low = yi - 1.96 * sei,
    ci_high = yi + 1.96 * sei,
    weight_percent = weights(res) / sum(weights(res)) * 100
  )

# Save meta-analysis data table
write.xlsx(
  list(
    Synthetic_Data = synthetic_data,
    Meta_Analysis_Data = meta_data,
    Meta_Analysis_Summary = data.frame(
      Metric = c("Number of studies", "Pooled log mean", "Lower 95% CI",
        "Upper 95% CI", "I2_percent", "tau2", "QE", "QEp"),
      Value = c(
        res$k,
        as.numeric(res$b),
        res$ci.lb,
        res$ci.ub,
        res$I2,
        res$tau2,
        res$QE,
        res$QEp
      )
    )
  ),
  file = "EnvChemMetaFlow_Meta_Analysis_Output.xlsx",
  overwrite = TRUE
)

# -----
# 5. Publication-style forest plot
# -----
png(
  filename = "EnvChemMetaFlow_Forest_Plot_32_Studies.png",
  width = 3000,
  height = 3600,
  res = 300
)

par(mar = c(5, 12, 4, 4))

forest(
  res,
  slab = meta_data$Study_Label,
  xlab = "Log Mean Concentration (ng/L)",
  mlab = "RE Model",
  psize = 1,
  cex = 0.65,
  header = FALSE
)

title("Synthetic Demonstration Forest Plot", cex.main = 1.1)

dev.off()

pdf(
  file = "EnvChemMetaFlow_Forest_Plot_32_Studies.pdf",
  width = 9,
  height = 11
)

par(mar = c(5, 12, 4, 4))

```

```

forest(
  res,
  slab = meta_data$Study_Label,
  xlab = "Log Mean Concentration (ng/L)",
  mlab = "RE Model",
  psize = 1,
  cex = 0.65,
  header = FALSE
)

title("Synthetic Demonstration Forest Plot", cex.main = 1.1)

dev.off()

# -----
# 6. Publication-style funnel plot
# -----
png(
  filename = "EnvChemMetaFlow_Funnel_Plot_32_Studies.png",
  width = 2400,
  height = 2000,
  res = 300
)

funnel(
  res,
  xlab = "Observed Outcome",
  ylab = "Standard Error",
  main = "Funnel Plot for Meta-Analysis",
  shade = c("white", "gray90", "gray80")
)

dev.off()

pdf(
  file = "EnvChemMetaFlow_Funnel_Plot_32_Studies.pdf",
  width = 8,
  height = 6.5
)

funnel(
  res,
  xlab = "Observed Outcome",
  ylab = "Standard Error",
  main = "Funnel Plot for Meta-Analysis",
  shade = c("white", "gray90", "gray80")
)

dev.off()

# -----
# 7. Egger's regression test
# -----
egger_test <- regtest(res, model = "rma")

capture.output(
  egger_test,
  file = "EnvChemMetaFlow_Egger_Test.txt"
)

# -----
# 8. Subgroup analysis by compound
# -----
res_compound <- rma(
  yi = yi,
  sei = sei,
  mods = ~ factor(Compound),
  data = meta_data,
  method = "REML"
)

capture.output(
  summary(res_compound),
  file = "EnvChemMetaFlow_Subgroup_Compound_MetaRegression.txt"
)

# -----

```

```

# 9. Subgroup analysis by water type
# -----
res_water <- rma(
  yi = yi,
  sei = sei,
  mods = ~ factor(WaterType),
  data = meta_data,
  method = "REML"
)

capture.output(
  summary(res_water),
  file = "EnvChemMetaFlow_Subgroup_WaterType_MetaRegression.txt"
)

# -----
# 10. Descriptive summary table
# -----
summary_table <- synthetic_data %>%
  group_by(Compound) %>%
  summarise(
    Studies = n(),
    Mean = round(mean(Mean_ng_L), 2),
    Median = round(median(Mean_ng_L), 2),
    Minimum = round(min(Mean_ng_L), 2),
    Maximum = round(max(Mean_ng_L), 2),
    .groups = "drop"
  )

write.xlsx(
  summary_table,
  file = "EnvChemMetaFlow_Descriptive_Summary.xlsx",
  overwrite = TRUE
)

# -----
# 11. End message
# -----
cat("EnvChemMetaFlow synthetic dataset and meta-analysis outputs generated successfully.\n")
cat("Generated files include dataset, forest plot, funnel plot, Egger test, subgroup analyses, and summary tables.\n")

# =====
# PART B. OPTIONAL FIGURES: PRISMA, HEATMAP, MODERATOR PLOT
# =====

# =====
# EnvChemMetaFlow Supplementary R Code: Optional Figures
# PRISMA flow diagram, pollutant distribution heatmap,
# and meta-regression moderator plot
# =====

# Required packages
required_packages <- c("dplyr", "ggplot2", "readxl", "openxlsx", "DiagrammeR", "metafor")

for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
}

library(dplyr)
library(ggplot2)
library(readxl)
library(openxlsx)
library(DiagrammeR)
library(metafor)

# -----
# 1. Load synthetic data
# -----
data <- read_excel("EnvChemMetaFlow_Synthetic_Dataset.xlsx")

data <- data %>%
  mutate(
    Log_Mean = log(Mean_ng_L),

```

```

    SE_Log_Mean = SD / (Mean_ng_L * sqrt(N)),
    Variance = SE_Log_Mean^2
  )

# -----
# 2. PRISMA-style flow diagram
# -----
prisma <- grViz("
digraph prisma {
  graph [layout = dot, rankdir = TB]
  node [shape = box, style = rounded, fontname = Helvetica, fontsize = 10]

  A [label = 'Records identified through database searching\\n(n = 420)']
  B [label = 'Records after duplicates removed\\n(n = 310)']
  C [label = 'Records screened by title and abstract\\n(n = 310)']
  D [label = 'Records excluded\\n(n = 220)']
  E [label = 'Full-text articles assessed for eligibility\\n(n = 90)']
  F [label = 'Full-text articles excluded\\n(n = 58)\\nNo extractable data = 24\\nWrong matrix = 16\\nReview only =
10\\nIncomplete statistics = 8']
  G [label = 'Studies eligible for synthetic template calibration\\n(n = 32)']
  H [label = 'Synthetic demonstration dataset generated\\n(n = 32 simulated studies)']
  I [label = 'Studies included in methodological meta-analysis demonstration\\n(n = 32)']

  A -> B -> C -> E -> G -> H -> I
  C -> D
  E -> F
}
")

# To export the PRISMA figure, run:
# DiagrammeRsvg::export_svg(prisma) |> charToRaw() |> rsvg::rsvg_png("EnvChemMetaFlow_PRISMA_Flow_Diagram.png")

# -----
# 3. Pollutant distribution heatmap
# -----
heatmap_data <- data %>%
  group_by(Compound, Country) %>%
  summarise(Mean_Concentration = mean(Mean_ng_L), .groups = "drop")

heatmap_plot <- ggplot(
  heatmap_data,
  aes(x = Country, y = Compound, fill = Mean_Concentration)
) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(Mean_Concentration, 0)), size = 3) +
  labs(
    title = "Synthetic Pollutant Distribution Heatmap",
    x = "Country",
    y = "Compound",
    fill = "Mean concentration (ng/L)"
  ) +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggsave(
  "EnvChemMetaFlow_Pollutant_Distribution_Heatmap.png",
  heatmap_plot,
  width = 9,
  height = 5.5,
  dpi = 300
)

ggsave(
  "EnvChemMetaFlow_Pollutant_Distribution_Heatmap.pdf",
  heatmap_plot,
  width = 9,
  height = 5.5
)

# -----
# 4. Meta-regression moderator analysis
# -----
res_water <- rma(
  yi = Log_Mean,
  sei = SE_Log_Mean,
  mods = ~ factor(WaterType),
  data = data,
  method = "REML"

```

```

)

capture.output(
  summary(res_water),
  file = "EnvChemMetaFlow_MetaRegression_WaterType_Output.txt"
)

moderator_summary <- data %>%
  group_by(WaterType) %>%
  summarise(
    Mean_Log = mean(Log_Mean),
    SD_Log = sd(Log_Mean),
    N = n(),
    SE = SD_Log / sqrt(N),
    CI_Lower = Mean_Log - 1.96 * SE,
    CI_Upper = Mean_Log + 1.96 * SE,
    .groups = "drop"
  )

metareg_plot <- ggplot(
  moderator_summary,
  aes(x = WaterType, y = Mean_Log)
) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = CI_Lower, ymax = CI_Upper), width = 0.15) +
  geom_text(aes(label = paste0("n=", N)), vjust = -1.0, size = 3.5) +
  labs(
    title = "Meta-regression Demonstration: Environmental Matrix as Moderator",
    x = "Moderator: Water Type",
    y = "Predicted log mean concentration (ng/L)"
  ) +
  theme_minimal(base_size = 12)

ggsave(
  "EnvChemMetaFlow_MetaRegression_Moderator_Plot.png",
  metareg_plot,
  width = 7.5,
  height = 5.5,
  dpi = 300
)

ggsave(
  "EnvChemMetaFlow_MetaRegression_Moderator_Plot.pdf",
  metareg_plot,
  width = 7.5,
  height = 5.5
)

cat("Optional EnvChemMetaFlow figures generated successfully.\n\n")

# =====
# PART C. SUBGROUP FOREST PLOT BY WATER TYPE
# =====

# =====
# EnvChemMetaFlow Supplementary R Script
# Subgroup Forest Plot by Water Type
# =====
# This script uses the same 32-study synthetic dataset generated for
# the EnvChemMetaFlow demonstration manuscript.
#
# Important:
# These data are synthetic and are used only for workflow demonstration.
# =====

required_packages <- c("readxl", "dplyr", "metafor", "openxlsx")

for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
}

library(readxl)
library(dplyr)
library(metafor)

```

```

library(openxlsx)

# -----
# 1. Load same 32-study synthetic dataset
# -----
data <- read_excel("EnvChemMetaFlow_Synthetic_Dataset.xlsx")

# -----
# 2. Prepare effect size
# -----
# Log mean concentration is used because the synthetic dataset contains
# both pharmaceutical micropollutants and nitrate, which differ greatly
# in concentration scale.

data <- data %>%
  mutate(
    Study_Label = paste(StudyID, Country, Compound, sep = " | "),
    Log_Mean = log(Mean_ng_L),
    SE_Log_Mean = SD / (Mean_ng_L * sqrt(N)),
    Variance = SE_Log_Mean^2
  )

# -----
# 3. Overall random-effects model
# -----
res_overall <- rma(
  yi = Log_Mean,
  sei = SE_Log_Mean,
  data = data,
  method = "REML"
)

summary(res_overall)

# -----
# 4. Subgroup/meta-regression model by water type
# -----
res_water <- rma(
  yi = Log_Mean,
  sei = SE_Log_Mean,
  mods = ~ factor(WaterType),
  data = data,
  method = "REML"
)

summary(res_water)

capture.output(
  summary(res_water),
  file = "EnvChemMetaFlow_Subgroup_MetaRegression_WaterType_Output.txt"
)

# -----
# 5. Subgroup summary statistics
# -----
subgroup_summary <- data %>%
  group_by(WaterType) %>%
  group_modify(~{
    m <- rma(
      yi = .x$Log_Mean,
      sei = .x$SE_Log_Mean,
      method = "REML"
    )
    data.frame(
      Studies_n = m$k,
      Pooled_Log_Mean = as.numeric(m$b),
      CI_Lower = m$ci.lb,
      CI_Upper = m$ci.ub,
      I2_percent = m$I2,
      tau2 = m$tau2,
      Q = m$QE,
      Q_pvalue = m$QEp
    )
  }) %>%
  ungroup()

overall_summary <- data.frame(
  WaterType = "Overall",

```



```

    Studies_n = res_overall$k,
    Pooled_Log_Mean = as.numeric(res_overall$b),
    CI_Lower = res_overall$ci.lb,
    CI_Upper = res_overall$ci.ub,
    I2_percent = res_overall$I2,
    tau2 = res_overall$tau2,
    Q = res_overall$QE,
    Q_pvalue = res_overall$QEp
  )

  subgroup_summary_all <- bind_rows(subgroup_summary, overall_summary)

  write.xlsx(
    subgroup_summary_all,
    file = "EnvChemMetaFlow_Subgroup_Analysis_WaterType.xlsx",
    overwrite = TRUE
  )

  # -----
  # 6. Publication-style subgroup forest plot
  # -----
  # The 'byvar' style can be generated manually by splitting the data by
  # WaterType and adding subgroup model summaries.

  data <- data %>%
    arrange(WaterType, Compound, StudyID)

  png(
    filename = "EnvChemMetaFlow_Subgroup_Forest_Plot_WaterType.png",
    width = 3000,
    height = 3900,
    res = 300
  )

  par(mar = c(5, 13, 4, 4))

  forest(
    res_overall,
    slab = data$Study_Label,
    order = order(data$WaterType, data$Compound, data$StudyID),
    xlab = "Log Mean Concentration (ng/L)",
    mlab = "Overall RE Model",
    cex = 0.60,
    psize = 1,
    header = FALSE
  )

  title("Subgroup Forest Plot by Water Type", cex.main = 1.1)

  dev.off()

  pdf(
    file = "EnvChemMetaFlow_Subgroup_Forest_Plot_WaterType.pdf",
    width = 9.5,
    height = 12
  )

  par(mar = c(5, 13, 4, 4))

  forest(
    res_overall,
    slab = data$Study_Label,
    order = order(data$WaterType, data$Compound, data$StudyID),
    xlab = "Log Mean Concentration (ng/L)",
    mlab = "Overall RE Model",
    cex = 0.60,
    psize = 1,
    header = FALSE
  )

  title("Subgroup Forest Plot by Water Type", cex.main = 1.1)

  dev.off()

  cat("Subgroup forest plot and subgroup summary table generated successfully.\n")

```