А.А. Кораблев, А.В. Пнюшков, А.В. Смирнов

ТЕХНОЛОГИЯ СОЗДАНИЯ БАЗ ОКЕАНОГРАФИЧЕСКИХ ДАННЫХ НА ПРИМЕРЕ СЕВЕРО-ЕВРОПЕЙСКОГО БАССЕЙНА АРКТИКИ

A.A. Korablyev, A.V. Pnyushkov, A.V. Smirnov

TECHNOLOGY OF COMPILING OCEANOGRAPHIC DATABASES: A CASE STUDY OF THE NORTH EUROPEAN ARCTIC BASIN

Статья посвящена технологии создания океанографических баз данных на примере Северо-Европейского бассейна. Обсуждается последовательность подготовки «очищенных» массивов данных от объединения начальных наборов данных до расчета объективно анализированных полей океанографических параметров. Особое внимание уделено контролю качества данных, методологии объективного анализа и функциональным возможностям разработанного программного обеспечения.

The article considers technology of compiling oceanographic database for the North European Basin. A sequence of compiling «clean» datasets is discussed from merging the original datasets to computing the objectively analyzed fields of oceanographic parameters. Special attention is given to data quality control, objective analysis methodology and functional capabilities of the developing software.

Введение

Значительно возросшие объемы океанографической информации, необходимой для решения оперативных и климатических задач, требуют построения информационно-вычислительных систем с использованием современных программных и технических средств ее хранения, обработки и представления. Разработка программного обеспечения, позволяющего осуществлять процедуры преобразований из различных форматов, контроля качества, интерполяции, объективного анализа, является необходимым условием эффективного усвоения и анализа данных. Достоверность итоговых наборов данных (НД) подразумевает выполнение целого ряда требований, в том числе полноту использованных начальных источников, обоснованность алгоритмов проверки качества, методов вертикальной интерполяции и объективного анализа, обеспеченность расчета средних характеристик. Важным аспектом при этом является оценка величин возможных ошибок (инструментальных наблюдений, интерполяции), на основании которых можно судить о правомерности использования НД для анализа.

Северо-Европейский бассейн (СЕБ), включающий Норвежское, Гренландское и Баренцево моря, занимает особое место в глобальной климатической системе. Он является транзитной зоной между Атлантическим океаном и

Арктическим бассейном, где происходит разнонаправленный перенос свойств, во многом определяющий климат умеренных и высоких широт [Малинин, Гордеева, 2003]. Межгодовые и долгопериодные изменения интенсивности процессов взаимодействия верхнего слоя океана с атмосферой, горизонтального и вертикального обмена в СЕБ оказывают прямое влияние на глобальную термохалинную циркуляцию [Dickson et al., 2002], адвекцию тепловых и соленостных аномалий в сопредельные районы [Karcher et al, 2003], погодные условия и климат северо-западной Европы и Арктики [Bengtsson et al., 2003]. Роль обратных связей, региональные аспекты перестройки циркуляции, вклад внутренней динамики в изменения климата и сравнительный анализ современного потепления в Арктике и потепления первой половины XX в., соотношение природной и антропогенной составляющих изменчивости, предсказание будущих изменений и их воздействие на биопродуктивность – далеко не полный перечень вопросов, решение которых невозможно без всестороннего анализа исходных данных и постоянного мониторинга океанического климата в СЕБ.

Цели и задачи исследований

Анализ источников океанографической информации для СЕБ показал, что на сегодняшний день не существует полного единого набора данных для этого региона. Проверка показала, что в базе World Ocean Database 2001 г. [Conkright et al., 2002] практически отсутствуют измерения за 90-е годы. Такая же ситуация наблюдается и в базе Арктического и Антарктического НИИ, созданной ранее [Ivanov, Korablev, 1996]. В то же время в 90-е годы в СЕБ в рамках различных международных проектов (ESOP, VEINS, TRACTOR, CONVECTION) получен значительный объем наблюдений. К примеру, институт морских исследований (IMR, Берген, Норвегия) проводит регулярные съемки с ежегодным числом выполненных станций, превышающим 2500. Для выполнения представленных исследований было использовано около 20 начальных НД, прошедших первичный контроль качества (рис. 1). В действительности их число было больше.

Целый ряд близких по составу НД были объединены между собой, например данные по кораблю погоды «М», поступившие из разных источников. Некоторые НД были замещены более новыми, например, вместо BARKODE и CLIMBAR использовался климатический атлас Северных морей 2004 г. [Matishov et al., 2004].

Перейдя к практическим аспектам создания информационно-вычислительной системы океанографических данных, сформулируем основные требования к ее построению. Прежде всего – это полнота использования исходных НД. Метаданные должны быть максимально полными и соответствовать стандартам, принятым в мировых океанографических центрах. Проверка качества информации должна осуществляться на основе стандартизованных ал-

горитмов, станционного и сравнительного контроля. Программное обеспечение должно обеспечивать оперативное пополнение баз данных и расчет объективно анализированных полей океанографических параметров в узлах заданной сетки.

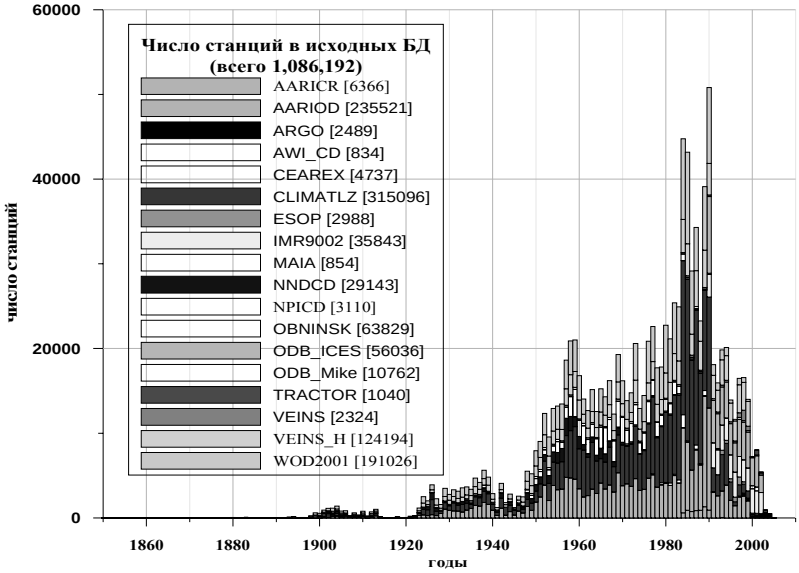

Рис. 1. Распределение количества станций по годам в исходных БД

Основные результаты исследований

Реализация перечисленных принципов предполагает многоступенчатый, последовательный процесс, описание которого и является предметом данной статьи (табл. 1). Прежде всего необходимо остановиться на программных средствах разработки и хранения данных. На основании анализа предлагаемых на рынке промышленных серверов хранения информации был выбран продукт компании Borland – Interbase. Это недорогая система управления базами данных (СУБД), полностью удовлетворяющая требованиям хранения имеющихся объемов океанографической информации.

СУБД Interbase интенсивно развивается, в настоящее время выпущена версия с номером 7,5.

Разработка метаданных и реляционных связей между таблицами – необходимый этап построения любой базы данных. В океанографии основной единицей хранения является океанографическая станция. В состав обязательных полей, на основе которых станция однозначно идентифицируется, входят координаты станции, дата и время ее выполнения. Однако время выполнения станции не всегда известно, особенно для исторических данных. Для сохранения уникальности набора перечисленных полей в метаданные добавлено еще одно поле – версия станции.

Таблица 1

Технологическая последовательность разработки океанографической базы данных Северо-Европейского бассейна

№	Последовательность выполнения процедур
1	Определение состава метаданных и связей между таблицами БД.
2	Создание сопровождающих баз данных рельефа дна (GEBCO), береговой черты, кодов стран и судов.
3	Разработка конверторов данных из начальных источников в форматы хранения СУБД Interbase и загрузка данных отдельные БД.
4	Обновление метаданных за счет добавления глубин дна из GEBCO и последнего измеренного горизонта на станции.
5	Удаление станций из БД, находящихся за пределами географической области СЕБ.
6	Контроль качества и установка флаг в качества на станцию и измеренные значения параметров (проверка диапазона изменчивости; проверка максимального допустимого числа горизонтов; проверка последовательности горизонтов; проверка дубликатов горизонтов; проверка на минимальное допустимое число горизонтов; контроль устойчивости и вертикальных градиентов; контроль на абсолютные дубликаты станций).
7	Слияние баз начальных источников с учетом типов инструментов и создание объединенных БД. Автоматический и экспертный контроль дублей, пополнение метаданных и профилей, создание полных единиц хранения.
8	Интерполяция на стандартные горизонты и типизация одноградусных квадратов на основании месячных распределений градиентов океанографических параметров средствами объективного анализа. Определение зон повышенной изменчивости (прибрежные, фронтальные, прикромочные зоны и области открытого океана).
9	Контроль качества и установка флагов качества на основании послойного анализа стандартных отклонений для отдельных месяцев за весь период наблюдений.
10	Интерполяция (линейная, Лагранж, Рейнигер-Росс) на стандартные горизонты с учетом флагов качества.
11	Повторная типизация одноградусных квадратов на стандартных горизонтах на основании месячных распределений градиентов океанографических параметров.
12	Контроль качества и установка флагов качества на основании анализа стандартных отклонений на стандартных горизонтах для отдельных месяцев за весь период наблюдений.
13	Объективный анализ океанографических полей и расчет набора месячных карт с заданным пространственным разрешением.
14	Расчет сезонных, среднемноголетних распределений параметров на основе месячных карт с учетом ошибки интерполяции. Расчет аномалий океанографических параметров с заданным временным и пространственным разрешением.

Исходя из требований нормализации, метаданные хранятся в двух связанных таблицах Station (15 полей) и Station_Info (10 полей). Остановимся на их составе. Помимо перечисленных таблица Station включает поле качества станции, название источника данных, судна, страны и четырех полей, характеризующих глубину места на станции. Первое из них соответствует измеренной глубине, три остальных (в точке станции, минимальная и максимальная в радиусе 5 км) добавлены из 1’ рельефа дна GEBCO [www.ngdc.noaa.gov/mgg/gebco]. Дополнительные глубины оказываются полезными при анализе качества метаданных, при измерении глубины и оценке принадлежности станции исследуемому району. Таблица Station_Info содержит международные коды страны и судна. В принципе эти коды должны однозначно идентифицировать судно, однако в ряде случаев это не

так. Указанные поля используются при загрузке данных (см. ниже) и присутствуют в большинстве начальных источников. Для устранения неоднозначности в кодах на этапе загрузки было проведено их объединение из разных источников (WOD2001, ICES, CLIMATL и др.) и создана единая таблица с указанием источника кода. Перечислим остальные поля в таблице Station_Info: номер станции в рейсе, код института, код проекта, код инструмента, уникальный номер из источника, название вторичного источника, идентификатор рейса. Уникальный номер из источника необходим для обратной связи с исходной базой данных. Код инструмента важен для оценки возможной инструментальной ошибки наблюдений. Всего используется восемь кодов, соответствующих основным типам измерительных датчиков в океанографии (неизвестный прибор, батометрия, CTD, STD, XCTD, MBT, XBT, попутные наблюдения).

Профили океанографических параметров хранятся в отдельных таблицах. Все таблицы однотипны и содержат по пять полей: абсолютный номер станции, последовательный номер горизонта, значение горизонта, значение параметра, флаг качества. Различие таблиц профилей заключается в формате хранения величины параметра. Внешний ключ, связывающий таблицы профилей с таблицами метаданных, позволяет осуществлять каскадное обновление и удаление данных в базах.

Исходя из стоящих задач, все разрабатываемое программное обеспечение было разделено на две части. Первое приложение (ODB3ALoad) предназначено для работы с исходными данными, начальной проверки качества, нахождения дублей, формирования объединенной базы данных, интерполяции на стандартные горизонты, контроля на стандартные отклонения и ряда других сервисных процедур. Итоговым продуктом приложения, после выполнения всей технологической цепочки преобразования данных является объединенная база океанографических данных. Второе приложение, о котором будет сказано ниже (ODB3A), представляет собой пользовательский интерфейс доступа к океанографической информации, или оболочку океанографической базы данных.

Общий вид пользовательского меню приложения ODB3ALoad представлен на рис. 2, последовательность пунктов которого, в целом, соответствует последовательности преобразования данных.

В пункте меню File реализована возможность выбора базы данных в режиме локального или удаленного доступа. В двух информационных строках, расположенных в нижней и верхней частях главной формы, отображается информация о названии и размере базы, количестве станций, географических границах их положения и периоде наблюдений. Пункт меню LoadData содержит конверторы из различных исходных форматов. Всего разработано более 15 различных конверторов, это отражает общую ситуацию с многообразием представления начальных данных в океанологии. Даже в случае использования однотипных форматов (ICES) детали форматов различны, что не позволяет использовать единый конвертор. Ряд сервисных процедур заключен в пункте меню Service. Наиболее универсальными из них являются процедуры

добавления глубин дна на станции, минимальной и максимальной глубины в радиусе 5 км из файла 1’ топографии и последнего измеренного на станции горизонта, определяемого на основании анализа всех имеющихся профилей.

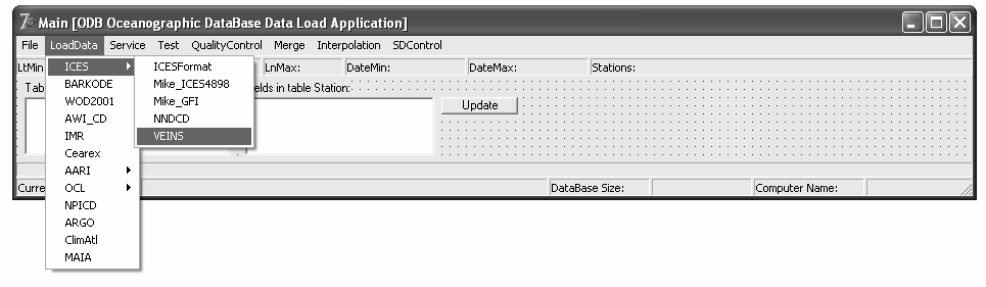

Рис. 2. Общий вид главной формы приложения ODB3ALoad с развернутым пунктом меню модулей конвертации исходных данных в формат Interbase.

Процедуры предварительного контроля качества сосредоточены в пункте меню QualityControl. Набор используемых алгоритмов определяет состав данных, который будет использоваться на всех последующих шагах. Всего проводится шесть проверок, которые приводят к изменению состава данных. Физически данные в большинстве случаев не удаляются, вместо этого устанавливаются флаги качества (табл. 2).

Таблица 2

Установка флагов качества (StFlag) в таблице STATION и на измеренные значения (Flag_) в таблицах океанографических параметров

№	Поле StFlag в таблице Station	Поле Flag_ в таблицах параметров
1	2	3
1	Наблюдения на станции ошибочны	Ошибочное значение
2	Наблюдения на станции подозрительны	Подозрительное значение
3	Гидрохимия не прошла контроль качества	Не используется
4	Метаданные пополнены из разных источников	Значение выходит за значение 5σ
5	Ряды наблюдений пополнены из разных источников	Значение выходит за значение 4σ
6	Не используется	Значение выходит за значение 3σ (квадрат, период)
7	NODC код судна не уникален или не найден в таблице ShipCode_List (SupportTables.ib).	Значение находится в высокоградиентной зоне
8	Разрешающая точность датчика солености < 0,02psu	Значение получено на основании анализа дубликатов
9	Разрешающая точность датчика температуры < 0,01°C	Значение не прошло проверку на вертикальную устойчивость (t, s) или на значение вертикального градиента
10	Дата станции изменена (формат времени)	Значение не прошло проверку на диапазон изменчивости
11	Дата станции изменена (число дней месяца)	Глубина измеренного горизонта пересчитана из давления в метры

1	2	3
12	Отсутствует время выполнения станции (нет часов)	Единицы измерения для гидрохимического параметра пересчитаны
13	Не используется	Интерполированное значение (по данным источника)
14	Время перехода между станциями одного рейса превышает допустимое значение	Ошибочное значение (по данным источника)
15	Станция на берегу	Подозрительное значение (по данным источника)
16	Последний измеренный горизонт на станции (по всем параметрам) выходит за диапазон минимального / максимального значения глубины GEBCO в радиусе 5 км от станции	Потеря точности измерения (по данным источника)

Флаги качества, устанавливаемые на станцию, формируются на основе анализа глубины последнего горизонта и глубины места, даты и времени выполнения станции, разрешающей способности датчиков температуры и солености, соответствия кодовым таблицам судов. Флаги качества для измеренных океанографических параметров устанавливаются как на этапе загрузки из начальных источников (если таковые имеются), так и на основании анализа, например, о пересчете давления в глубину, преобразовании гидрохимических параметров, выхода значений за интервалы стандартных отклонений.

Контроль качества начинается с проверки соответствия измеренных значений параметра интервалу его физической изменчивости. Значения, выходящие за диапазон, удалялись. Контроль горизонтов осуществляется по нескольким алгоритмам, включающим разрежение подробных CTD профилей с дискретностью 5 м в слое до 400 м и 10 м – глубже. Первый и последний горизонты сохраняются, как в исходном профиле. В дальнейшем проводится анализ последовательности и дублирования горизонтов. Процедура контроля на минимально допустимое число наблюдений на станции предназначена для увеличения однородности конечного набора данных. Станция сохраняется для дальнейшего анализа, если число горизонтов наблюдений по температуре или солености не менее трех. Это достаточно жесткое условие приводит к отсечению значительного числа поверхностных измерений и некоторого числа прибрежных станций. Однако, использование таких данных для последующего анализа может внести значительное снижение точности оценки средних значений параметров и их дисперсий [Emery & Thomson, 2003]. Контроль устойчивости проводился по диапазонам инверсии плотности. Допустимым отрицательным градиентом плотности для слоя 0–30 м является значение – 0,03 кг/м^4; для слоя 30 – 400 м – (–0,02 кг/м^4) [Conkright et al., 2002]. Глубже 400 м данные с отрицательным значением градиента плотности помечались соответствующим флагом.

Одной из проблем объединения данных из различных начальных источников, на которой необходимо остановиться особо, является наличие большо-

го количества дубликатов. Очевидно, что оценить полноту и точность представления метаданных и профилей на станции можно только на основании сравнения всех имеющихся вариантов. Фактически, для получения полной единицы хранения информации (океанографической станции) требуется полное взаимное сравнение с сохранением только измеренных величин. Программные модули создания исходных баз данных (БД) и нахождения дубликатов находятся в пункте меню Merge (см. рис. 2). Вначале создается объединенная база данных, структура которой отличается от стандартной. Добавлено несколько дополнительных таблиц, в том числе, содержащих все метаданные из начальных баз. На практике для сокращения числа одновременно анализируемых станций создается набор БД, содержащих станции за определенный временной интервал (как правило, год). Профили параметров физически остаются в исходных БД, обращение к которым осуществляется по названию источника и абсолютному номеру.

После создания объединенной БД проводится анализ дубликатов. Алгоритм состоит из последовательного выполнения процедур их идентификации. На каждом шаге находятся все станции с одинаковой датой и координатами с учетом возможной ошибки округления (+/– 1’), после чего начинается автоматический анализ на возможное дублирование. Используется следующая типизация в соответствии с последовательностью контроля: абсолютные дубли (совпадают все метаданные и профили), полные дубли (совпадают основные метаданные и профили температуры, солености и кислорода), TSO2 дубли (метаданные могут отличаться, профили температуры, солености и кислорода совпадают), разреженные CTD профили, интерполированные варианты профилей, многосуточные станции (метаданные могут совпадать, профили разные). Нахождение интерполированных вариантов профилей проводится путем перекрестной интерполяции значений одного профиля на горизонты второго и определения профиля невязки. Если отличие составляет менее 0,01 ºС для температуры, 0,01 psu для солености и 0,05 мл/л для кислорода, профили считаются интерполированными вариантами.

Все эти процедуры выполняются автоматически, каждый алгоритм присваивает свой флаг, на основании которого станция пишется в новую базу. На каждом шаге сравнения станций происходит пополнение метаданных и профилей. В результате отсутствующие метаданные и ряды, принадлежащие одной станции, переходят к другой. В связи с этим первоначальное распределение начальных источников по порядку имеет значение, так как пополнение метаданными происходит каскадно, т.е. от исходных баз с менее высоким приоритетом к базам с более высоким приоритетом. Таким образом, формируется полная единица хранения на основе всей имеющейся совокупности станций. Если хотя бы одна из выбранных на данном шаге станций не была идентифицирована, проводится экспертный контроль.

В качестве примера эффективности автоматического нахождения дублей приведем пример подготовки объединенной базы данных для июня 1990 г. Число начальных источников составляло 10 с общим количеством станций 7889. С помощью рассмотренных выше алгоритмов было автоматически идентифицировано более 95% станций, из них абсолютных дублей 643 (записано в базу)/1846 (не пишется в базу), полных дублей 53/157, TSO2 дублей 162/1846, интерполированных/разреженных профилей 46/624, многосуточных станций 1534. На основании экспертного контроля в базу добавлено 115 станций, признано дублями 251. Итоговый, очищенный от дублей массив данных составил 39% от исходного, со станциями из всех 10 начальных источников. Это еще раз подтверждает отсутствие единого достаточно полного НД для СЕБ.

Процедуре вертикальной интерполяции профилей океанографических параметров предшествует анализ распределений градиентов, типизация их структуры и установка флагов качества на основе расчета стандартных отклонений. Это необходимо для оптимального выбора интервалов стандартных отклонений. Как правило, зоны повышенных градиентов приурочены к прибрежным районам, фронтальным зонам и прикромочным областям. Игнорирование повышенной изменчивости в таких районах может привести к необоснованному исключению данных. Алгоритм типизации одноградусных квадратов основан на результатах расчетов градиентов океанологических параметров, полученных после объективного анализа месячных данных за весь период наблюдений.

Для расчета интервалов стандартных отклонений был разработан отдельный программный модуль. Предусмотрено два варианта анализа: послойный (наблюдения в окрестности стандартных горизонтов) и на стандартных горизонтах. Вначале проводится пошаговое объединение станций, выбранных для каждого месяца года. Помесячный анализ выбран не случайно, так как месячные поля являются базовыми для объективного анализа и расчета на их основе различных средних значений. На первом шаге выбирается крайняя северо-западная станция во всей совокупности. Начальный радиус поиска соседних станций задается равным 50 км. Выбор начального и предельного радиуса поиска определяется статистическими свойствами распределений океанографических параметров и радиусами пространственной корреляции. Анализ характеристик вариограмм показал, что предельно допустимые расстояния, на которых сохраняются значимые значения корреляций для СЕБ, составляет 180–400 км в зависимости от географического положения и глубины. На следующем шаге анализируется число выбранных станций в слое или на стандартном горизонте. Если число станций в 50-километровой области меньше предельно допустимого значения, происходит пошаговое увеличение радиуса поиска до предельного. Выбор предельного числа станций, удовлетворяющих требованию получения несмещенной оценки, определяется статистическими свойствами совокупности. На практике это значение задавалось не меньше 21. Далее,

для выбранных станций в заданном слое рассчитываются стандартные отклонения (σ) и устанавливаются флаги качества в соответствии с попаданием измеренного значения за пределы интервалов изменчивости 3, 4 или 5 σ. В зависимости от принадлежности станции к градиентной зоне или области открытого океана (на основании типизации одноградусных квадратов) устанавливается соответствующий флаг на измеренное значение в таблицах профилей. В дальнейшем совместный анализ этого флага и флагов статистической проверки, позволяет накладывать ограничения на использование океанографических параметров.

Для проведения вертикальной интерполяции каждый профиль предварительно очищается с использованием установленных ранее флагов качества и контроля на стандартные отклонения. Далее проводится интерполяция с использованием одного из трех методов: линейной интерполяции, интерполяции Лагранжа или метода Рейнигера – Росса [Reiniger and Ross, 1968]. Линейная интерполяция проводится по двум горизонтам, для последних двух типов необходимы соответственно 3 или 4 уровня наблюдений. Интерполяция возможна при выполнении ряда условий взаимного расположения горизонтов, это же относится и к выбору метода интерполяции. Выбор критериев выполнения интерполяции может сильно повлиять на результирующие вертикальные профили, поэтому требует осторожного подхода. В отличие от линейной интерполяции, методы Лагранжа и Рейнигера – Росса позволяют получать более гладкие профили, но возможна и генерация локальных экстремумов.

В подходе, примененном в данной работе, имеется ряд существенных отличий от методов, используемых при подготовке атласа WOD2001 [Conkright et al., 2002]. Во-первых, задан более жесткий критерий проведения интерполяции между двумя горизонтами. При этом используется линейная функция изменения критерия с глубиной, в отличие от послойной. Максимальная разница глубин двух горизонтов на поверхности не может превышать 5 м, на горизонте 3500 м – 1000 м. Использование линейной функции позволяет избежать скачков на границах слоев, что физически более оправдано. Во-вторых, аналогичные условия используются при интерполяции по методу Рейнигера – Росса по четырем горизонтам. Ограничения на глубины взаимного расположения «ближней» и «дальней» пары точек также задаются линейной функцией. И последнее существенное отличие: задан дополнительный критерий использования нелинейных методов. Если полученное на стандартном горизонте значение более чем на 20% отличается от значения, полученного с помощью линейной интерполяции, используется ее линейный вариант, что позволяет избежать появления необоснованных экстремумов.

Работы по загрузке и проверке данных весьма трудоемкие. Достаточно сказать, что для выполнения всех описанных процедур к настоящему моменту было отлажено 45 программных модулей, а общее число строчек программного кода превысило 20 000.

Рассмотрим кратко второе приложение (ODB3A) океанографической базы, которое в настоящее время находится в стадии разработки, – пользовательский интерфейс доступа к БД. Основными требованиями к любой современной оболочке являются возможность ее использования неподготовленным пользователем, наличие необходимых функциональных возможностей и представления данных в удобной графической и табличной формах. Общий вид главного меню приложения представлен на рис. 3.

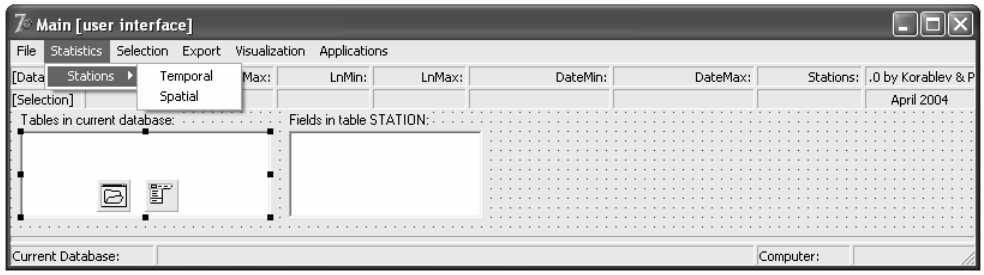

Рис. 3. Общий вид главной формы приложения ODB3A.

В отличие от загрузочного модуля в пункте меню File, добавлены функции, позволяющие получать справочную информацию о содержании всех баз данных, находящихся в текущем каталоге. Распределение числа станций в начальных источниках на рис. 1 получено с помощью именно этой процедуры. Возможность выборки станций реализована в пункте Selection. При формировании запроса можно использовать большую часть полей, содержащихся в таблицах метаданных, включая координаты, время выполнения, название страны, судна, глубины дна, соотношения глубины дна и последнего горизонта и др. Кроме того, возможен выбор станций вдоль заданного разреза и вокруг заданной точки с указанием радиуса поиска. Полученные в результате запроса станции записываются в отдельную таблицу и могут быть использованы во всех программных модулях приложения. Результаты выборки представляются в табличном виде и в виде карты на фоне береговой черты. Выборки могут сохраняться и загружаться с диска.

Справочная информация о станциях в БД может быть получена в пункте меню Statistics. Имеется доступ к двум модулям – распределению по времени и по пространству. Полный набор статистических характеристик по времени включает вывод информации о составе таблиц профилей (число станций и измерений), распределении числа станций по годам, месяцам, параметрам, странам, судам. Распределение числа измерений по вертикали для каждого из параметров также доступны в графическом и табличном вариантах. Пространственное распределение станций в БД дается в виде карты с измененяемой цветовой гаммой и возможностью задания числа градаций по количеству станций в квадратах. Возможна навигация по карте с интерактивным выводом информации о квадрате, над которым позиционирован курсор.

Предусмотрена возможность выгрузки станций в тестовом формате по заданному району, времени и набору профилей (пункт Export). В отличие от встроенных в средства работы с Interbase программ (IBConsole, IBExpert), модуль позволяет выгружать данные в виде связанных по абсолютному номеру таблиц БД.

Графическое представление профилей океанографических параметров реализовано в пункте Visualization. Пункт меню «Applications» зарезервирован для программных модулей, реализующих различные методы стандартных вычислений, принятых в океанографии. На данный момент доступны модули расчета временных диаграмм распределений океанографических параметров и их аномалий по глубине для заданного района и гистограмм распределения числа станций в заданном слое. В ближайшие время предполагается значительно усовершенствовать функциональные методы оболочки базы данных, в том числе за счет добавления новых методов обработки данных.

В заключение раздела остановимся на сравнительном анализе разрабатываемого ПО с мировыми аналогами в области хранения и обработки океанологической информации, такими, как «Ocean Data View» [Schlitzer, 2002], «Java Ocean Atlas» [http://odf.ucsd.edu/joa/jsindex.html], «Climatic Atlas of the Arctic Seas 2004» [Matishov et al., 2004]. Перечисленные продукты предоставляют пользователю разнообразные возможности доступа и визуализации данных в определенном формате. Несомненным лидером является разработанная в AWI (Бременхафен, Германия) система доступа к данным, позволяющая использовать встроенные НД с топографией, характеристиками ледового покрова, географическими названиями объектов, стоком рек, донными осадками и некоторыми спутниковыми наблюдениями. Оболочка Климатического Атласа Северных морей предназначена для визуализации данных, находящихся на диске файлов в формате html. Функциональные назначение «Java Ocean Atlas» ограничивается предоставлением пользователю графического интерфейса визуализации информации, хранящейся во внутреннем формате, с добавлением возможностей по экспорту данных.

В разрабатываемой ПО (ODB) используется промышленная система управления базами данных, что позволяет обращаться к данным на языке структурированных запросов (SQL) и создавать распределенные системы, работающие в многопользовательском режиме. Использование объектно-ориентированного языка (Object Pascal) дает возможность создавать интерактивные приложения любого уровня сложности. Важным отличием является наличие модулей конвертации данных из большого числа исходных форматов и возможность добавления новых конверторов. Среди наиболее существенных функциональных особенностей ODB можно выделить:

– встроенные процедуры проверки качества информации;

– возможность объединения исходных НД с автоматическим и экспертным определением дубликатов и созданием полных единиц хранения;

– наличие системы флагов качества, на основании которой можно фильтровать используемые данные;

– возможность расчета объективно-анализированных полей с любым пространственно – временным разрешением, в том числе на нерегулярной сетке станций.

Объективный анализ океанографических полей

Одним из важных направлений в технологической цепочке создания климатических баз данных является задача восстановления информации в узлах регулярной сетки. Теоретическая основа использованного для этих целей метода объективного анализа (ОА) заключается в минимизации функционала ошибки при варьировании набора весов интерполяции по сравнению с любым другим подобным набором [Cressie, 1993; Isaaks, Srivastava, 1989]. После определения весовых коэффициентов, удовлетворяющих условию минимума невязки, значение интерполируемой величины может быть найдено как линейная комбинация весовых коэффициентов и самих значений параметра. Использование такой процедуры для океанографических данных дает ряд преимуществ по сравнению с другими методами, часто используемыми для получения гидрометеорологических полей в узлах регулярной сетки. В качестве таковых можно упомянуть:

– метод ближайшего окружения, при котором весовые коэффициенты точек, попадающих внутрь радиуса поиска, считаются постоянными, а само значение в узле вычисляется по формуле

$$Z^* = \sum_i \omega_i Z_i \ , \ \omega_i = \text{const}, \qquad (1)$$

где Z^* – значение параметра в интерполируемом узле; Z_i – значение параметра в точке с известными координатами; ω_i – вес i-го наблюдения,

– метод обратных расстояний;

– метод линейной интерполяции внутри триангуляционной сетки;

– метод полиномиальной аппроксимации и ряд других методов.

Применение объективного анализа на основе методов кригинга позволяет провести процедуру интерполяции с данными, обладающими рядом специфических особенностей. К наиболее важным из них можно отнести:

– пространственную неоднородность, т.е. наличие зон либо с полным отсутствием наблюдений, либо с их малым количеством;

– существенную анизотропию, под которой понимается отличие свойств статистической взаимосвязи данных в различных направлениях. Этот феномен может быть вызван, к примеру, наличием фронтальных зон и т.п.;

– наличие в данных пространственных и временных трендов, связанных с изменяющимися во времени и пространстве параметрами физических процессов.

Одной из базовых составных частей метода объективного анализа является оценка степени взаимосвязи исходных данных. Такую оценку можно получить

из анализа ковариационной функции. Зачастую, вместо нее рассматривается вариограмма, которая также отражает меру пространственной взаимосвязи интерполируемой величины [Cressie, 1993]. Для нахождения значений экспериментальной вариограммы $\gamma(d)$ можно воспользоваться формулой (2):

$$\gamma(d) = \frac{\sum_{N_p} [Z(x+d) - Z(x)]^2}{2\,N_p}, \qquad (2)$$

где $Z(x)$ – значение параметра в точке x; $Z(x + d)$ – значение параметра на расстояние d от точки x; N_p – количество пар точек; d – расстояние между точками наблюдений.

Связь между значениями ковариационной функции $C(d)$ и эмпирической вариограммой определяется следующей зависимостью:

$$\gamma(d) = C(0) - C(d). \qquad (3)$$

Результат проведения объективной интерполяции во многом зависит от вида теоретической кривой, описывающей экспериментальную вариограмму. Использование для анализа ковариационной функции в этом плане представляется более затруднительным из-за ограничений, которые вводятся на класс аппроксимирующих функций, главное из которых – это положительная определенность [Cressie, 1993].

В качестве теоретической модели вариограммы могут использоваться следующие функции:

1. Линейная зависимость $\gamma(d) = a_0 + b\,d$, когда величина $\gamma(d)$ линейно возрастает с расстоянием. Очевидно, что такая функция не является ограниченной сверху и фактически содержит внутри себя неисключенную трендовую составляющую; a_0 и b – константы, определяемые эмпирически.

2. Экспоненциальная модель: $\gamma(d) = c_0 + c_s\left[1 - \exp(d/a)\right]$; c_0, c_s, a – эмпирические коэффициенты.

3. Сферическая модель: $\gamma(d) = c_0 + c_s\left[\frac{3}{2}\frac{d}{a} - \frac{1}{2}\left(\frac{d}{a}\right)^3\right]$ при $d < a$

$$\gamma(d) = c_0 + c_s \text{ при } d \geq a.$$

Следует упомянуть, что параметр a, входящий в экспоненциальную и сферическую модели, зачастую интерпретируется как радиус пространственной корреляции данных.

Рассмотрим чуть более подробно сферическую и экспоненциальную модели как наиболее часто употребляемые в гидрометеорологии. Анализ расчетов показал, что экспоненциальная модель дает более быстрое уменьшение весовых коэффициентов внутри радиуса корреляции по сравнению со сферической, что

особенно заметно на малых расстояниях. Для линейного масштаба порядка радиуса корреляции значения, полученные по этим моделям, получаются более близкими, а начиная с расстояния, равного *a,* практически одинаковыми.

Одно из исходных предположений (допущений) при проведении процедуры объективного восстановления данных – это отсутствие в рядах наблюдений трендовой составляющей [Cressie, 1993; Isaaks, Srivastava, 1989]. Для этого из исходных данных вычитается их среднее значение, и дальнейшие действия производятся с вычисленными аномалиями. После определения всех весовых коэффициентов оптимальной интерполяции аномалии складываются с учетом их веса и прибавляются к фоновому значению. Поэтому в случае, когда интерполируемый узел расположен далеко от точек наблюдения, вычисленный параметр в этом узле будет стремиться к своему фоновому значению. На практике условие стационарности анализируемого процесса зачастую не выполняется, что приводит к необходимости введения в методику ОА ряда дополнительных процедур, которые позволяют обойти это ограничение.

Отдельного внимания в этом плане заслуживает реализованный в модуле ОА метод декомпозиции вариограммы. Применение композитной модели позволяет разделить общую изменчивость моделируемого процесса на характерные диапазоны крупно- и мелкомасштабной динамики. Фильтруя ту или иную составляющую, в зависимости от целей исследования, можно получить детальное описание поля интерполируемой величины с исключенными крупномасштабными или мелкомасштабными вариациями. Для применения такого подхода были реализованы два различных метода:

1. Из всей исходной информации вычитался линейный двумерный тренд, ассоциирующийся с крупномасштабной изменчивостью. Вычисленные характеристики крупномасштабных процессов использовались для оценки теоретической модели их вариограммы. Параметры вариограммы для меньших масштабов определялись по значениям разницы между непосредственными данными измерений и соответствующих им значений трендовой составляющей.
2. Второй метод основан на применении декомпозиции непосредственно к самой эмпирической вариограмме.

Этот подход частично совпадает с так называемым медианно очищенным кригингом, описанным в работах [Isaaks, Srivastava, 1989; Raty, Gilbert., 1998].

Перейдем непосредственно к алгоритмам, реализованным в модуле ОА. Традиционно в геостатистике различаются несколько видов кригинга:

1. Простой кригинг (ПК) (simple kriging).
2. Ординарный кригинг (ОК) (ordinary kriging).
3. Универсальный кригинг (УК) (universal kriging).

Процедура простого кригинга предполагает устойчивость в оценке среднего значения домена, внутри которого проводится интерполяция, т.е. отсутствие систематической ошибки в его определении. Для этого типа объективного анализа не выдвигается жесткого условия равенства единице суммы ве-

совых коэффициентов. Система линейных уравнений ПК для нахождения весов интерполяции λ_β может быть представлена уравнением (4)

$$\sum_\beta \lambda_\beta \gamma_{\alpha\beta} = \gamma_{\alpha 0}\,, \tag{4}$$

где $\gamma_{\alpha\beta}$ – значение вариограммы для пар точек измерений; $\gamma_{\alpha 0}$ – значение вариограммы для узла интерполяции; β – количество измерений.

Значение в сеточном узле определяется как линейная комбинация наблюдений Z_β (5):

$$Z^* = \sum_\beta \lambda_\beta Z_\beta + \lambda_0\,, \tag{5}$$

где $\lambda_0 = m\left(1 - \sum_\beta \lambda_\beta\right) = m\lambda_m$; m – среднее значение внутри домена.

При этом средняя квадратическая ошибка интерполяции σ^2 [ε] будет определяться формулой (6):

$$\sigma^2[\varepsilon] = C(0) - \sum_\beta \lambda_\beta C_{\beta 0}\,. \tag{6}$$

В случае, когда расположение узлов с данными приводит к корректной процедуре интерполяции, последнее слагаемое в правой части уравнения (5) стремится к нулю. Отклонения суммы весовых коэффициентов от единицы может служить формальным критерием качества проведения анализа в конкретном узле. Напрямую с этим связана и такая величина, как средняя квадратическая ошибка интерполяции, определяемая как доля общей дисперсии поля. Исходя из принципов построения системы кригинга, квадратическая ошибка будет определяться только взаимным расположением точек с исходной информацией относительно узла регулярной сетки и относительно друг друга. Как и следует ожидать, в случае, когда все данные расположены далеко от интерполируемого узла, погрешность метода будет фактически равна средней квадратической ошибке внутри домена.

Недостатков, связанных с изменениями среднего значения внутри области интерполяции, лишен метод ординарного кригинга, одним из условий которого является равенство единице суммы весовых коэффициентов. Математическая формулировка метода ОК представлена в виде системы уравнений: (7) – (10):

$$\sum_\beta \lambda_\beta \gamma_{\alpha\beta} + \mu = \gamma_{\alpha 0}\,, \tag{7}$$

$$\sum_\beta \lambda_\beta = 1\,, \tag{8}$$

$$Z^* = \sum_{\beta} \lambda_\beta Z_\beta \,, \tag{9}$$

$$\sigma^2[\varepsilon] = C(0) - \sum_{\beta} \lambda_\beta C_{\beta 0} - \mu \,. \tag{10}$$

Для соблюдения условия (8) в исходную систему весов кригинга вводится дополнительный параметр μ, который в литературе принято обозначать как параметр Гаусса [Cressie, 1993; Isaaks, Srivastava, 1989].

Введение дополнительного ограничения на выбор параметров весовых коэффициентов можно интерпретировать как сужение области поиска оптимальных коэффициентов с β-мерного пространства до поиска в пространстве с размерностью β-1, удовлетворяющем условиям (8). Строго говоря, полученное решение уже не будет оптимальным с точки зрения минимизации ошибки интерполяции как в случае простого кригинга, поскольку минимум в пространстве поиска решения будет больше либо равен общему минимуму ошибки.

Как уже было упомянуто выше, для исследуемого процесса выдвигается требование стационарности [Cressie, 1993; Isaaks, Srivastava, 1989]. В случае, когда в исходных данных наблюдается очевидный тренд, – на графике вариограммы это выглядит как возрастающая функция расстояния, можно применить метод универсального кригинга. Этот метод наряду с весовыми коэффициентами интерполяции учитывает наличие в данных пространственного тренда. Такой подход может быть оправдан в случае, когда интерполяция проводится для больших областей с явно выраженной закономерностью в распределении вдоль какого-то направления. Например, для температуры воды, когда априори известна тенденция ее повышения при продвижении от полюсов к низким широтам.

Однако применение модели кригинга с линейным или нелинейным трендом может привести к серьезным ошибкам, особенно в областях с редкой сетью наблюдений. Вычисленные параметры тренда по отдельным измерениям могут быть проэкстраполированы на расстояния, сопоставимые с масштабом домена и тем самым дать абсолютно нереалистичные значения интерполируемого параметра. Вероятность больших искажений, вносимых процедурой универсального кригинга, а также тот факт, что ряд районов СЕБ остается фактически непокрытым сетью наблюдений, заставляет отказаться от его реализации в модуле ОА климатических полей региона.

В итоге можно сформулировать иерархию перечисленных методов восстановления. Наиболее просто реализуемым методом, безусловно, является метод простого кригинга, следующим по сложности является ординарный кригинг, вводящий ряд ограничений на весовые коэффициенты интерполяции. И наиболее сложный тип – это универсальный кригинг, учитывающий наличие в исходных данных неисключенного пространственного тренда. Как можно заметить, ординарный кригинг является частным случаем универсального кригинга при отсутствии в данных крупномасштабной трендовой составляющей.

Проведенные серии экспериментов с программным модулем объективного анализа показали, что наиболее корректные результаты получаются при использовании сферической и экспоненциальной моделей вариограммы для процедуры ординарного кригинга. При проведении сравнительного анализа использовался его точечный вариант. Отличие точечного кригинга от также реализованной в модуле блочной модели заключается в том, что в результате оценивается значение величины в заданной точке пространства, а не внутри выбранного блока. Применение процедуры блочного кригинга, как показывает практика, приводит к определенного рода сглаживанию интерполируемого поля с масштабом пространственного размера ячейки. Для реализации такого подхода достаточно в правую часть матрицы весов (7) подставить среднее интегральное значение вариограммы для всех расстояний от точки наблюдений до множества точек домена.

На практике для оценки среднего значения вариограммы достаточно использовать разбиение домена сеткой 4×4 узла. Таким образом, можно подвести некоторый итог изложенному выше и перечислить основные результаты, достигнутые в этом направлении:

– разработан модуль объективного анализа океанологических полей с использованием прямого доступа к БД на основе Borland Interbase 7,0;

– реализованы методы простого и ординарного кригинга, позволяющие корректно анализировать нестационарные поля;

– реализован метод декомпозиции статистической структурной функции, дающий возможность выделять крупно- и мелкомасштабную изменчивость в исследуемых полях;

– реализованные алгоритмы позволяют оценить погрешность восстановления информации в каждом узле интерполяции, что позволяет судить о ее качестве.

В качестве иллюстрации результатов объективного анализа океанографических полей, полученных по данным объединенной БД для СЕБ за июнь 1990 г., приведем поля температуры и солености (рис. 4, 5) на горизонте 50 м.

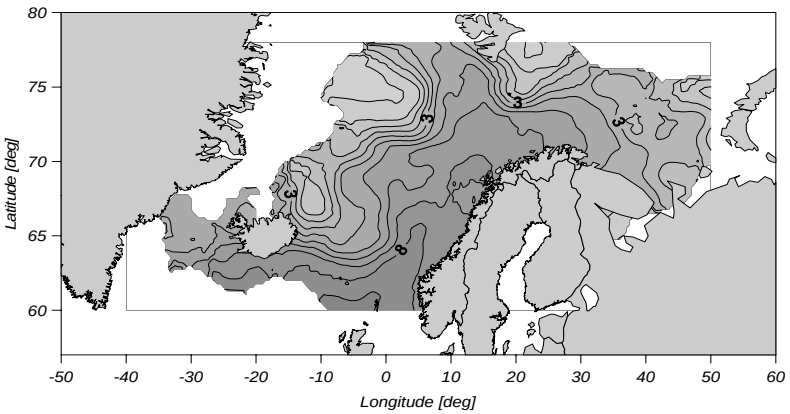

Рис. 4. Распределение температуры воды в июне 1990 г. на горизонте 50 м по данным объективного анализа

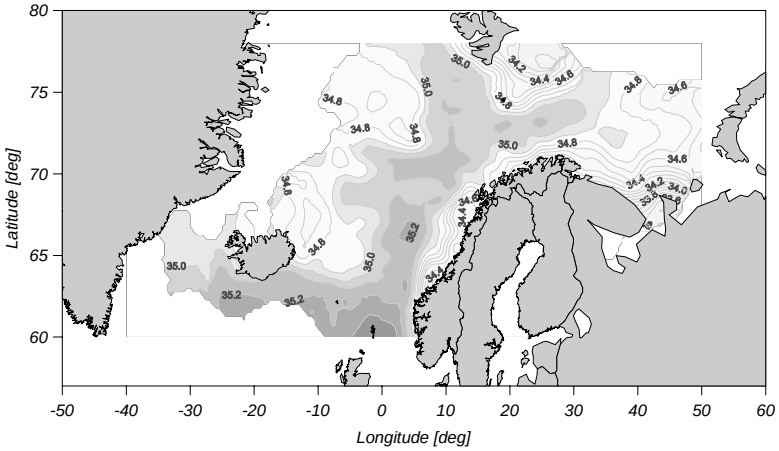

Рис. 5. Распределение солености воды в июне 1990 г. на горизонте 50 м по данным объективного анализа

Заключение и выводы

Рассмотренная последовательность преобразования данных представляет собой законченную технологическую цепочку подготовки объективно анализированных полей океанографических данных. Система разрабатывалась не только как средство для хранения и визуализации океанографических данных, но и как инструмент создания прошедших контроль качества БД, на основе первичных источников информации. Прозрачность и документированность всех промежуточных процедур является необходимым требованием для оценки применимости итоговых БД для научного анализа. Система флагов качества позволяет определять ограничения на использование данных. Очевидно, что задание более жестких критериев отбора данных приводит к формированию более сглаженных наборов и, следовательно, потери части естественной изменчивости. Определение ошибки, связанной с недостаточной обеспеченностью расчета средних значений, является ключевым моментом для оценки достоверности полученных распределений океанических параметров и расчета их аномалий. В качестве метода объективного анализа выбрана система кригинга, позволяющая получать оценки средних величин параметров в узлах заданной сетки и величину ошибки интерполяции.

Использование большого числа начальных источников океанографических данных для Северо-Европейского бассейна Арктики позволяет утверждать, что созданная база является наиболее полной. Пополнение метаданных и рядов на основе автоматического и экспертного контроля дает возможность сформировать наиболее полный состав хранения станций на основе сравнения дублирующей информации из разных источников.

Разработанное программное обеспечение является инструментом оперативного пополнения информации в океанографической базе данных, а модуль-

ный принцип его построения позволяет добавлять новые расчетные методы. Приложения могут быть использованы для создания наборов данных для задач оперативного обеспечения и усвоения в математических моделях по любому району Мирового океана.

Литература

1. *Малинин В.Н., Гордеева С.М.* Физико-статистический метод прогноза океанологических характеристик (на примере Северо-Европейского бассейна). – Мурманск: Изд-во ПИНРО, 2003. – 164 с.
2. *Bengtsson L., V.A. Semenov, O.M. Johannessen* The early Twentieth-Century warming in the Arctic – A possible mechanism // Journal of Climate, 2004, vol. 17, 4045–4057.
3. *Conkright M.E., Antonov J.I., Baranova O., Boyer T.P., Garcia H.E., Gelfeld R., Johnson D., Locarnini R.A., Murphy P.P., O'Brien T.D., Smolyar I., Stephens C.* World Ocean Database 2001, Volume 1: Introduction. Ed: Sydney Levitus, NOAA Atlas NESDIS 42, U.S. Government Printing Office, Washington, D.C., 2002. – 167 p.
4. *Cressie N.* Statistics for Spatial Data, revised edition, Wiley, New-York, 1993. – 928 p.
5. *Dickson B., Yashayaev I., Meincke, Turrell, Dye, Holfort*. Rapid freshening of the deep North Atlantic Ocean over the past four decades. // Nature, 2002, 416, 832–837.
6. *Emery W.J., Thomson R.E*. Data analysis methods in physical oceanography. Elsevier., Second and revised edition, Amsterdam-Tokyo, 2003, p. 638.
7. *Isaaks E.H., Srivastava R.M.* An Introduction to Applied Geostatistics, Oxford University Press, 1989. – 768 p.
8. *Ivanov V., Korablev A., Myakoshin O.* PC-adapted oceanographic database for studying climate shaping ocean processes. In Oceanology International 96. The Global Ocean – Towards Operational Oceanography, Conference Proceedings, vol. 1, UK, 1996, p. 89–99.
9. *Karcher M., Gerdes R., Kauker F., Köberle C.* Arctic warming – evolution and spreading of the 1990s warm event in the Nordic Seas and the Arctic Ocean. J. Geophys. Res., 2003, 108 (C2), 3034, doi: 10.1029/2001JC001265.
10. *Matishov, G., Zuyev A., Golubev V. et al.* Climatic Atlas of the Arctic Seas 2004: Part I. Database of the Barents, Kara, Laptev, and White Seas – Oceanography and Marine Biology. U.S. Government Printing Office, Washington D.C., NOAA Atlas NESDIS 58, 2004.
11. *Raty L., Gilbert M.* Large-scale versus small-scale variation decomposition, followed by kriging based on a relative variogram, in presence of a non-stationary residual variance // Geographic Information and Decision Analysis, 1998, vol. 2, № 2, p. 91–115.
12. *Reiniger R.F., Ross C.K.* A method of interpolation with application to oceanographic data. Deep Sea Res., vol. 15, 1968, p. 185–193.
13. *Shiltzer R.* Interactive analysis and visualization of geoscience data with Ocean Data View // Computers & Geosciences 28, 2002, 1211–1218.

.